Vol. 69, No. 1 January 1962

Psychological Review

Richard L. Solomon, Editor
University of Pennsylvania

CONTENTS

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

The *Psychological Review* is devoted to articles of theoretical significance to any area of scientific endeavor in psychology. Except for occasional articles solicited by the Editor, manuscripts exceeding 7,500 words will not be accepted. Ordinarily manuscripts which consist primarily of original reports of research should be submitted to other journals. However, reports of original research may be included in a major theoretical article which attempts to integrate several related original studies.

Because of the large number of manuscripts submitted, there is an inevitable publication lag of several months. Authors may avoid this delay if they are prepared to pay the costs of publishing their own articles; the appearance of articles by other contributors is not thereby delayed.

Manuscripts must be accompanied by an abstract of 100 to 120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstract may be obtained from the Editor or from the APA Central Office.

Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors should retain a copy of the manuscript, in case there is loss in the mail. Manuscripts should be addressed to the Editor, Richard L. Solomon, Psychology Department, College Hall, University of Pennsylvania, Philadelphia 4, Pennsylvania.

Helen Orr
Acting Production Manager

Published Bimonthly by the
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Prince and Lemon Streets, Lancaster, Pa.
1333 Sixteenth St. N.W., Washington 6, D. C.

$10.00 volume $2.00 issue

Second-class postage paid at Lancaster, Pa.

Send all communications, including address changes, to 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must arrive by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

VOL. 69, No. 1 JANUARY 1962

PSYCHOLOGICAL REVIEW

VERTICAL AND HORIZONTAL PROCESSES IN PROBLEM SOLVING [1]

HOWARD H. KENDLER
New York University

AND

TRACY S. KENDLER
Barnard College

The present paper is concerned with *an* approach—and not *the* approach—to the universally appealing but nevertheless unpopular research area of problem solving. Problems of problem solving have proved to be particularly refractory to psychologists. More often than not the uncommon researcher with the temerity to attack some aspect of reasoning retreats to more secure and conventional problems when he discovers that his sorties fail to achieve any impressive victory. As a result the literature of problem solving is almost chaotic because it is so heavily sprinkled with isolated bits of information (Duncan, 1959).

Perhaps the present stage of development of psychology does not justify the strategy of investigating such a complex phenomenon. Fortunately, or not, science has no built-in traffic lights to inform investigators when to proceed. It may be a risky and potentially unfruitful gambit to investigate problem solving but then again it may not be. In addition to the intrinsic interest of the area it does offer a challenge to those psychologists who are interested in testing the generality of any set of theoretical principles stemming from other areas of behavior (e.g., learning, perception).

This paper initially will make fleeting references to some methodological problems with which a researcher in the field of reasoning must contend. Then a simple pretheoretical model of problem solving will be described, followed by a report of research which the model generated, and which in turn is shaping the model itself.

METHODOLOGICAL PROBLEMS IN PROBLEM SOLVING RESEARCH

Anybody who does research is—or should be—aware that every decision he makes cannot be justified by facts or logic. Some decisions must be made on the basis of personal intuition. This is particularly true for the researcher in problem solving who must make three strategic decisions which cannot help but have profound influences on his research and the ideas they generate (Kendler, 1961). These decisions, which are not completely independent, are related to the place of problem solving in psychology, the use of complex or simple experimental

[1] An earlier version of this paper was delivered by Howard H. Kendler as an invited address to the 1960 meeting of the Eastern Psychological Association, which was held in New York City. The authors are indebted to the Office of Naval Research and the National Science Foundation for their support of the research reported in this paper.

tasks, and the selection of a pretheoretical model to guide research. Considering the volitional nature of these problems, as well as the current status of psychological knowledge, it would be both inappropriate and erroneous to consider these methodological problems as offering only one sensible alternative. Adopting this point of view would do much to minimize the needless disputation that seems to perennially surround matters of research strategy.

Accepting the principle that a basic research strategy is not simply an outgrowth of logical and factual considerations does not reduce one to making decisions in either a haphazard or random manner. A given strategy can be adopted on the basis of rational considerations as long as it is realized that other reasonable attitudes might lead to the adoption of different decisions.

The history of problem solving in particular and psychology in general suggests that problem solving can best be conceptualized not as a basic psychological process, but instead as one that reflects the interaction of more fundamental processes (e.g., learning, perception, and motivation).

If problem solving is not viewed as a unitary process, how is an appropriate experimental situation selected to investigate it? One possibility is that a problem can be selected from a "true life" situation such as troubleshooting electronic equipment. Or problems can be invented (Duncker, 1945; Maier, 1930) that capture the flavor, if only partially, of problems we meet in everyday life.

A more analytical approach can be taken to the selection of an experimental situation to investigate problem solving. If problem solving is compounded of elementary behavioral processes, then it may be more strategic to devise some simple problems in which the relationships of fundamental psychological mechanisms to problem solving are highlighted. That is, tasks should be devised not to duplicate or imitate everyday problems, but instead to isolate and magnify the basic mechanisms that operate in such complex tasks.

This analytical approach which is favored by the authors suffers from one major drawback. How is it possible to know the basic mechanisms of problem solving prior to their discovery? Obviously, excepting divination, there is no method. But this does not prevent the analytical approach from operating. The researcher can prejudge theoretical issues by formulating a model of what he guesses problem solving to be like. The model can guide the investigator in selecting the hypotheses to test, as well as the experimental situations in which to test them.

This brings us to the third and most important decision a problem solving researcher has to make: his choice of a pretheoretical model (Koch, 1959). A pretheoretical model is not equivalent to a theory. The criterion of validity cannot properly be applied to it because essentially a pretheoretical model is an informal conception that operates as an analogy (Lachman, 1960). It is conceivable that different models (e.g., learning, perception, information theory) can all lead to fruitful and valid theories of problem solving.

Psychologists have many possibilities from which to choose their model. These models can be conveniently divided into two main categories: the empirical model that springs primarily from experimental data, and the formal model that is usually generated by mathematical or logical systems. Among the empirical models that have achieved some acceptance are those

that are based on introspective findings (e.g., the four successive stage model of "preparation," "incubation," "inspiration," and finally "verification"), the facts of perception, and those of learning. Some formal models used are those dependent upon stochastic models, game theory, and the operation of computers.

The present authors adopted an S-R learning pretheoretical model. The decision no doubt was influenced by professional training and past research efforts. But other considerations entered. For the past 4 decades S-R learning psychologists have probably been the most active experimental and theoretical group in psychology. To some, if not a large, extent this can be attributed to the fruitful and cleansing effect S-R language has upon designing, reporting, and interpreting research. S-R language forces the psychologist to focus his attention on objectively defined environmental and behavior variables and thus encourages the collection of data and the testing of ideas. The efforts of S-R learning psychologists have supplied a host of facts, concepts, and hypotheses that can be exploited in an exploratory excursion into the realm of problem solving.

The facts and theories of learning, however, do not spontaneously coalesce to form a model that can guide research in problem solving. Some selection must be made. S-R learning theory does not represent a single organized formulation. Anyone who is familiar with the systematic orientations of Hull (1952), Guthrie (1952), Spence (1956), and Skinner (1953) is aware of this. Many of these systematic differences, however, become attenuated and some even disappear when viewed from the distance of problem solving behavior. It is possible and perhaps even profitable to develop a learning model for problem solving that ignores many of the points of disagreement among S-R theories.

Much of the objection to S-R language stems from the apparent discrepancy between active, flowing behavior and the inert, static, single S-R association. Using S-R language does not mean that complex behavior *actually* consists of S-R connections. After analyzing the concept of light, Toulmin (1953), concludes: "We do not *find* light atomized into individual rays: we *represent* it as consisting of such rays" (p. 29). Applying the same idea to the concept of the S-R association: "We do not *find* behavior atomized into individual S-R associations: we *represent* it as consisting of such S-R associations." The concept of the S-R association, therefore, must be judged not in terms of its ability to provide a clear image of behavior, but rather in its capacity to represent the facts of behavior.

PRETHEORETICAL MODEL OF PROBLEM SOLVING

An S-R model needs to represent two important characteristics of problem solving behavior. These characteristics are behavior is continuous, and at any one time behavior consists of several habits. The terms "horizontal" and "vertical" are used to refer to these processes; horizontal to the continuity of behavior against the dimension of time, and vertical to the assumption that independent levels of behavior (i.e., S-R units) occur simultaneously.

The assumption that S-R associations do not occur in isolation, but instead are linked together to form integrated, continuous behavior goes back many years (e.g., Watson, 1913). Today the process is most commonly referred to as chaining. Skinner (1953) and his associates have developed powerful techniques that shape

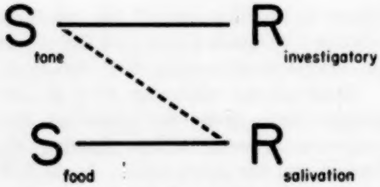

FIG. 1. An S-R representation of classical conditioning.

behavior into long, complicated chains. The mass of data they have collected suggests important principles governing habit chaining. There is litle doubt that when their quasitheoretical system is exploited fully with autoinstructional devices that important insights into problem solving behavior will emerge, particularly in relation to how an added bit of knowledge can trigger problem solution. The kind of chaining with which the Skinnerians have dealt (i.e., adding new S-R units to an already functioning chain) does not exhaust all the problems associated with the horizontal processes of problem solving. Of particular importance to problem solving is the *spontaneous* integration of separate habits which occurs when an organism infers the consequences of combining previously independent S-R units. This kind of chaining was investigated in Kohler's (1925) classical studies of insight and in the more controlled reasoning experiments of Maier (1930). More recently the authors (Kendler & Kendler, 1956, 1961; Kendler, Kendler, Pliskoff, & D'Amato, 1958) have tried to identify some of the important variables that enable children to combine separate experiences in order to solve an inference-type problem. Much of the research reported in this paper will be concerned with how mediated stimulus and response events aid in the formation of problem solving chains.

The assumption of vertical processes, i.e., the organism responds several different ways at any one time, is also not a novel one. Every psychologist is aware that organisms make several different responses simultaneously, although typically only one is attended to. Sometimes the different responses are interrelated, as is the case between the heart and respiration rates of a fearful organism. In other cases the different responses are independent, e.g., a person's conversation is uninfluenced by his tugging at his ear lobe. The best laboratory example of vertical processes, and one that has much relevance to problem solving, is shown in Figure 1. Those familiar with introductory psychology textbooks will recognize this diagram as representing classical conditioning. Notice that the two solid lines indicate independent S-R units which are operating simultaneously. One is the tone that initiates the "investigatory" response, and the other is the food which elicits salivation. Initially these two associations operate in a *parallel* fashion, but as a result of their simultaneous occurrence an *interaction* takes place which is expressed by the broken line representing the acquired conditioned response.

Obviously the brief reference to horizontal and vertical processes in which it is assumed fundamental S-R principles operate (e.g., discrimination, generalization, etc.) presents at best the barest skeleton of a model of problem solving. It needs the flesh and skin of experimental facts to give it solidity and theoretical principles to clothe it in scientific respectability. Let us now review some of the progress that has been made in this direction.

CONCEPT LEARNING AND UTILIZATION

Although the primitive model just described fails to generate any research by itself, it does suggest that individual experiments cannot be directed at

problem solving in its entirety. There are too many aspects to this phenomenon. The researcher, in designing an experiment, must scan the entire problem solving process and then focus upon that segment that promises to yield fruitful results and is also amenable to investigation.

For reasons that will become evident, it was decided to compare reversal and nonreversal shifts in a simple concept learning task. Figure 2 characterizes each kind of shift by showing a *simplified* version of an experimental situation used with children. The stimuli (cups) for their first discrimination differed simultaneously on two dimensions (size and brightness). The subject is rewarded for responses to one dimension (e.g., large cup is positive, small cup is negative). The other dimension is irrelevant. After learning the first discrimination, the subject is forced to shift to another response. In a reversal shift the subject is required to respond to the same dimension on which he was originally trained, but his overt choice has to be reversed, e.g., he has to shift from a *large* cup to a *small* one. For a nonreversal shift the previously irrelevant dimension becomes relevant, e.g., black becomes positive after large had been positive.

Buss (1953) reported that college students executed a reversal shift more rapidly than a nonreversal shift. He attributed this superiority to the intermittent reinforcements that retard the progress of a nonreversal shift. For example, in Figure 2,[2] when a subject is making a nonreversal shift from large positive to black positive, he is reinforced when choosing the large black cup in preference to the small white cup. This fortuitous reinforcement of the choice of the large cup helps maintain the size discrimination and hence retards the learning of the brightness discrimination. The reversal shift group, on the other hand, receives no reinforcement of the previously correct responses, since they are 100% *non*reinforced.

This analysis is at best incomplete. The work of Kendler and Vineberg (1954) suggested that adult human concept learning cannot be represented adequately by a single-unit S-R theory in which the external stimulus is directly connected to the overt response. Instead, a mediational mechanism (see Figure 3) is required which assumes that the external stimulus evokes an implicit response which produces an implicit cue that is connected to the overt response.

It would be useful to digress for a

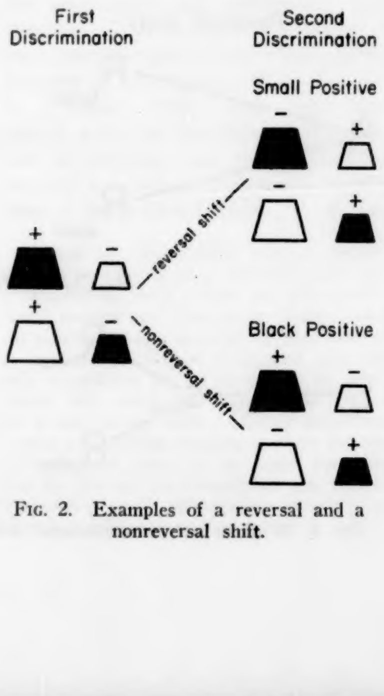

FIG. 2. Examples of a reversal and a nonreversal shift.

[2] The purpose of Figure 2 is to clarify the meaning of both a reversal and nonreversal shift. It would be misleading to believe that it represents *exactly* the methodology of "reversal-nonreversal" studies reported in this paper. For all experiments reported, except that of Buss (1953), designs were used that controlled for fortuitous intermittent reinforcements effects in a nonreversal shift.

$$S \text{ — } \boxed{r \text{-----} s} \text{ — } R$$

FIG. 3. A schematic representation of the mediational hypothesis.

moment to comment about the epistemological status of these inferred stimulus and response events which are enclosed in the rectangle to emphasize their hypothetical character. Although not directly observable, they are "tied to" environmental and behavioral events. The basic assumption of the mediational hypothesis, at least for the time being, is that the implicit stimulus and response events obey the same principles that operate in observable S-R relationships.

The mediational hypothesis has generated confusion. Perhaps the following brief statements will clarify some possible areas of misunderstanding.

1. The mediational hypothesis is neither new nor revoluntionary. Meyer (1911) and Watson (1913) referred to it, and Hull (1930) gave it a more formal status by coining the concept of the "pure stimulus act." Guthrie (1952) has always laid heavy stress on a mediational-type hypothesis when emphasizing the importance of proprioceptive stimulation in learning.

2. The implicit stimulus and response events *need not* be conceived as having an existence independent of their relation to independent and dependent variables. These implicit events are theoretical constructs. Their epistemological status is closer to such concepts as drive and habit than to directly observable stimulus and response events.

Some mediating events can conceivably and probably will be coordinated to introspective reports, language behavior, muscular movements, and other observable events. Coordinations of this sort can be useful in developing mediational theory. But such coordinations are not *essential* to mediational theory. The fact that genes are not

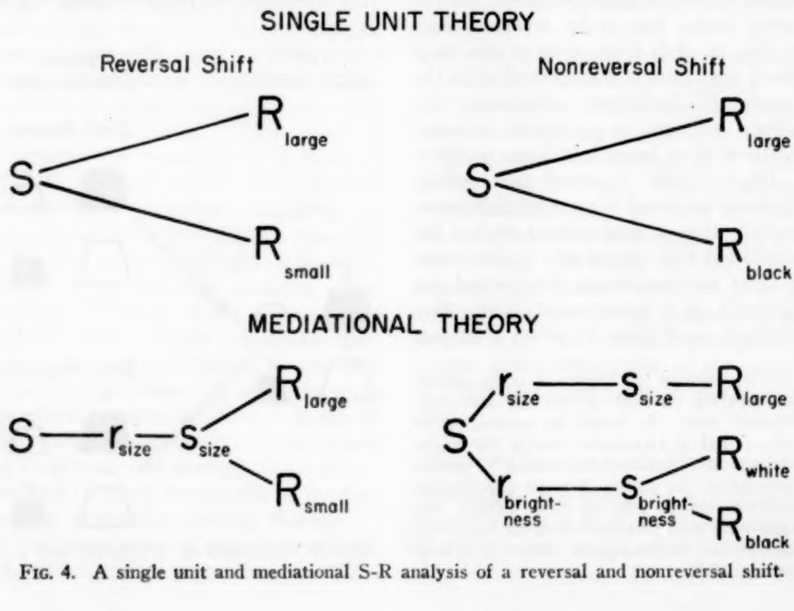

FIG. 4. A single unit and mediational S-R analysis of a reversal and nonreversal shift.

directly observable (at least according to the geneticists consulted) does not interfere with their theoretical and practical usefulness. Even if it were possible to observe a gene directly, it would be necessary to distinguish between it as an observable entity and as a concept within a nomological network. It would be unwise, and strategically shortsighted, to *identify* mediational events with introspective reports or language behavior, or other observable events. The "validity" of the mediational mechanism does not depend on being coordinated with observable events, but depends instead on being utilized in a successful explanatory system.

Figure 4 characterizes reversal and nonreversal shifts in terms of both a single unit S-R analysis and a mediational one.[3] It would be predicted, according to a single unit hypothesis, that if fortuitous intermittent reinforcements were eliminated from a nonreversal shift, it would occur more rapidly than a reversal shift. The reason for this is that at the time of the shift the difference between the strength of the dominant incorrect habit and the to-be-correct habit is much greater for the reversal, as compared to the nonreversal shift. Consequently more training will be required to make the correct habit dominant in a reversal shift. According to the mediational theory the situation is entirely different. A reversal shift enables the subject to utilize the same mediated response. Only the overt response has to be changed. A nonreversal shift, on the other hand, required the acquisition of a *new* mediated response, the cues of which have to be attached to a *new* overt response. Because the old mediational sequence has to be discarded and a new one formed, the nonreversal shift should be executed more slowly than a reversal shift.[4] Thus, if it were possible to eliminate fortuitous intermittent reinforcements, then the stage would be set for a crucial experiment testing the conflicting implications of the single-unit and mediational S-R theories. The results of a series of such crucial experiments (Buss, 1956; Harrow & Friedman, 1958; Kendler & D'Amato, 1955) have been consistent with the mediational formulation in showing that college students execute a reversal shift more rapidly than a nonreversal shift. It is im-

[3] Figure 4 highlights the problem of what are the effective stimuli that are associated to the overt response in both a reversal and nonreversal shift. It is not intended to be a detailed analysis of which there may be several alternatives. For example, in a single unit theory the habit to choose the large container might result from learning two separate specific habits (e.g., the choice of a large black container when coupled with a small white one and the selection of a large white container when paired with a small black one). Another possibility, which would be consistent with Spence's theory (1936), is that the response is to the effective stimulus *large* since responses to the other features of the environment are not consistently reinforced. Similarly adult subjects in a reversal shift might use the mediator *size* or *large* or both. The effective stimulus which is controlling the organism's response must be determined by experimentation. The point made here is that the general implications of the single unit and mediational theories, as discussed in this paper, would be the same for a number of different effective stimuli.

[4] There are two possible ways of analyzing the superiority of a reversal shift over a nonreversal shift within an S–R mediational framework. One is to simply count the number of new associations that have to be formed. As Figure 4 indicates only one new association has to be formed in a reversal shift while two have to be formed for a nonreversal shift. Another possibility is that a mediating response is more difficult to extinguish than is an overt response. For the present the formulation can remain open-ended until information relevant to these two alternatives is gathered.

TABLE 1

MEAN NUMBER OF TRIALS TO CRITERION ON TEST DISCRIMINATION FOR SUBJECTS SCORING ABOVE AND BELOW THE MEDIAN ON THE TRAINING DISCRIMINATION

Group	Performance on training discrimination	
	Above Median (slow learners)	Below Median (fast learners)
Reversal	24.4	6.0
Nonreversal	9.0	15.5

portant to note that in a similar kind of problem rats find a nonreversal shift easier than a reversal shift (Kelleher, 1956). Thus, one is forced to conclude that a single unit S-R theory accurately represents the behavior of rats, while mediational S-R theory is required for the concept learning of articulate humans.

The discontinuity between the behavior of rats and college students directs one's attention toward the conditions responsible for the development of mediational processes. Somewhere on a hypothetical evolutionary dimension between the rat and college student there should be a point where a transition is made from a single unit to mediational control. An obvious place to locate this point would be in the behavior of young children.

A study with kindergarten children (Kendler & Kendler, 1959) showed that these children as a group executed a reversal and nonreversal shift at approximately the same rate. One might conclude that the point in human development was discovered which was psychologically halfway between the white rat and the college student, since the kindergarten children were neither responding in a single unit nor mediational manner, but instead in some compromise fashion. Another possibility is that the children had reached a transitional stage in development, in which the task to which they were subjected led some to function on a single unit basis, and others to operate with a mediational mechanism. If half of the subjects respond in each way, the total results would have revealed no difference between the two kinds of shifts.

The second alternative seems to fit the data. When the kindergarten children were divided into fast and slow learners on the basis of their performance in the first problem (training discrimination), slow learners performed during the second problem (test discrimination) according to the single unit theory; like rats they found a nonreversal shift easier. Fast learners, on the other hand, performed in accordance with the mediational theory; like college students, they found a reversal shift easier. These results were interpreted as demonstrating that these kindergartners, taken as a group, were in the process of developing mediating responses relevant to this task, and that some were further along than others.

If this interpretation be correct, then it would follow that for a group of younger (i.e., preschool) children a still smaller proportion should develop appropriate mediating responses. It

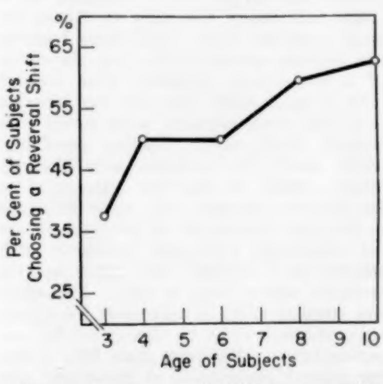

FIG. 5. Percentage of children responding in a reversal shift manner as a function of age.

would be expected that such a group, taken as a whole, would show clearcut evidence of the superiority of a nonreversal over a reversal shift. An experiment (Kendler, Kendler, & Wells, 1960) designed to test this hypothesis produced results consistent with this prediction; like rats, nursery school children found a nonreversal shift to be easier than a reversal shift.

In a very recent study the experimental procedure was modified so that after learning the initial discrimination, the children of 3, 4, 6, 8, and 10 years of age who served as subjects, had a choice of either responding in a reversal or a nonreversal manner. Under such circumstances, it would be expected that the proportion of children who respond in a reversal manner would increase with age. Figure 5 shows that the percentage of children who chose a reversal shift rose gradually from 37.5 at 3 to 62.5 at 10.

Generalizing from all of these results, it would seem that in their early development, children tend to respond in a manner consistent with a single unit S-R theory. With age, they develop a tendency to respond in a mediational manner. The last study cited suggests that it is, or will soon be, possible to ascertain the lawful relationship governing the course of this development.

The point of these experiments is not to classify children into one of two categories: rat-like or human-like. Their aim is to lay the groundwork for experiments designed to investigate the mediational process itself. If one wants to investigate mediational processes, does it not seem sensible to scrutinize them at the time when they are developing? Answering this question in the affirmative, it was decided to investigate the relationship between the hypothesized mediational processes and verbal behavior—a relationship everybody assumes to be intimate and important.

Particularly relevant to this attempt to coordinate verbalization with mediation were observations that during the course of the experiments just described, it was not uncommon for children to verbalize spontaneously the correct solution while simultaneously making an incorrect choice. A few children did this for many consecutive trials. This observation is relevant to the concept of vertical processes. Two chains of habits are occurring simultaneously. One has to do with verbal response; the other with the overt choice. For these children the two chains are parallel, that is, they do not interact.

Luria (1957), the Russian psychologist, made somewhat similar observations in his research with children. He explains this sort of phenomenon in the following way:

> In the early stages of child development, speech is only a means of communication with adults and other children. . . . Subsequently it becomes also a means whereby he organizes his own experience and regulates his own actions. So the child's activity is mediated through words (p. 116).

These observations and their interpretations of noninteracting parallel processes point to the complex interrelationships existing between verbal behavior on the one hand and problem solving on the other. If nothing else, they destroy the illusion that it is reasonable to describe an organism as verbal or nonverbal without considering the problem with which it is confronted. The terms verbal and nonverbal become meaningful—and fruitful—when related to specific problem solving tasks.

It would seem fruitful to investigate the cue function of words for children of two age levels. One possibility is that age influences problem solving

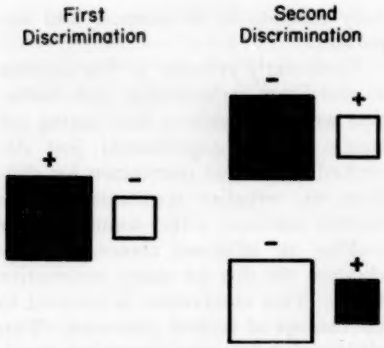

FIG. 6. The experimental procedure used to study the influence of verbal habits on a reversal shift.

only in so far as it leads to the acquisition of words. If younger children, say 4 years of age, could acquire the same words as 7-year-olds, they would solve a simple concept-learning problem the same way. The other possibility is that the acquisition of the verbal label by itself is not sufficient; the word must be integrated with other behavioral chains to influence problem solving behavior. And for this to happen some developmental changes must first take place.

In order to test these two alternatives, children of 4 and 7 years of age were presented with another variation of the reversal shift problem as shown in Figure 6. They initially learned a simple discrimination between a pair of stimuli that varied simultaneously in size and brightness. In the illustration provided in Figure 6, the large black square is correct. While they were learning, the children were required to verbalize aloud the stimuli to which they were responding. One-third learned to say "large" (or "small" as the case may be) by the simple device of instructing them to tell the experimenter which was correct, the large or the small one. Another third learned to say "black" (or "white") in a corresponding way. The remaining third was not required to say anything. After learning the discrimination, all subjects were presented with a reversal shift. In the example depicted in Figure 6, the shift is to small regardless of size. Thus, the group that initially described the correct stimulus as "large" had verbalized the relevant dimension. The verbal response of "black" was irrelevant to this reversal shift.

Figure 7 shows the results of the three experimental groups for the two age levels. If developmental processes affect the utilization of verbal responses in problem solving, then it would be expected that the three verbalization conditions (which produced a significant main effect) would influence the behavior of the two age groups differently. These results suggest, but not quite at a significant level, that there is an interaction effect. Figure 7 shows that the younger children profited by making the kind of verbal response appropriate to a reversal shift, while they were hindered by learning

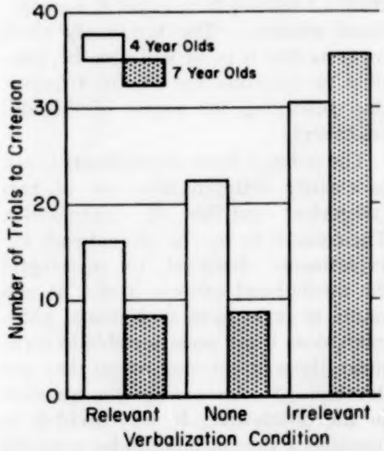

FIG. 7. The effect of verbalizations on a reversal shift for 4- and 7-year-old children.

inappropriate verbal responses. With no verbalization the 7-year-old children who presumably were responding largely in a mediational manner, accomplished a reversal shift much more rapidly than their younger counterparts. But unlike the 4-year-olds, they did not profit from being trained to make the relevant responses. At 7 years of age they are capable of making the response themselves and outside help appears to be of little use. In contrast, the influence of irrelevant verbalizations is marked. The performance of the 7-year-olds was even poorer than that of the 4-year-olds, suggesting that the interfering effects of being given an inappropriate mediated response are greater when one is capable of spontaneously generating the correct one (7-year-olds) than when one is not (4-year-olds).

How are these data to be explained? Attributing differences to developmental factors is not sufficient. It is necessary to represent developmental differences in terms of the concepts of the behavior model that is being used. That is, if a verbal label for a young child does not possess the same cue function as it does for an older child, then it becomes necessary to specify how and why this comes about. To some extent this has been done by emphasizing the transition from a single unit to a mediational system, as well as suggesting that with age an increase occurs in interaction among chains of different vertical levels. But obviously this analysis of the developmental process demands further theoretical and empirical development.

These studies are intimately related to the oft-reported finding that many species of subhuman animals are able to make a fairly rapid reversal shift *if* they receive a previous series of such shifts. Rats (Buytendijk, 1930; Krechevsky, 1932) show a marked improvement in executing successive reversals. They finally reach a point (Dufort, Guttman, & Kimble, 1954), in a T maze, where they learn to go to a new rewarded goal after making only one error. Even more dramatic are the rapid discrimination reversals exhibited by Harlow's (1949) monkeys. But fish (Wodinsky & Bitterman, 1957) exhibit only a slight improvement in successive reversals, while isopods (invertebrates) show no improvement (Thompson, 1957).

Because of the necessity to use somewhat different experimental procedures for different species, it is difficult to draw an unqualified conclusion about the ability of different species to transfer what has been learned from previous reversal shifts to a new one. But the suggestion is strong that as you ascend the evolutionary scale organisms acquire a greater capacity to generate cues that enable them to make rapid reversal shifts. This behavior, according to our analysis, borders on the language responses of humans. The main difference is that our human subjects, except those of a very young age, exhibit rapid reversals without any previous reversal training. Whereas the human automatically seems to generate a mediated response that provides the basis for his rapid reversal, the animal subject must gradually acquire an ability to respond appropriately to some response produced cue resulting from nonreinforcement of a previously correct response.

Up to now, the reversal and nonreversal technique has been used to investigate mediational and developmental variables. It has proved sufficiently flexible to be used in a study (Kendler, Glucksberg, & Keston, 1961) which was designed to lengthen a problem solving chain so that the interaction between various segments could be observed. In this study a

S	—— r ⇝	S	—— r ········	s	—— R
ready signal	orienting response	figure	symbolic response	symbolic cue	key pressing

FIG. 8. The hypothesized behavioral chain operating at the time the subject was being shifted to the second concept. (Capital letters refer to directly observable stimulus and response events, while small letters refer to those that are inferred.)

perceptual orienting S-R unit was added on to the mediational chain already described. Figure 8 illustrates in an oversimplified manner the behavioral sequence involved in this study in which subjects had to learn to press the correct button when two physically discrete and spatially separate stimulus patterns were projected on a screen at such a rapid rate that only one could be perceived on any trial. During the learning of each of two successive concepts (involving either a reversal or nonreversal shift), the subject had to pay attention to the relevant stimulus pattern while ignoring the irrelevant one. Thus, in order to make the correct overt response consistently, a subject initially had to make the appropriate orienting response in order to perceive the relevant stimulus pattern to which he had to make the correct mediational response which served as the cue for the key-pressing act.

An experimental design was used in which, at the time of the shift from the first to the second concept, one group had already learned the appropriate orienting response as well as the appropriate mediating act. They needed only to learn a new terminal key-pressing response. The shift, for them, was easy to make. In contrast, the behavior of three other experimental groups was significantly worse. One group had to learn a new orienting response, e.g., look to the left instead of the right. Another group had to learn a new mediated response (i.e., they were required to make a nonreversal shift). The last group had to acquire both a new orienting and mediated response. The fact that the groups which were missing one or both of the necessary behavior units (orienting and mediated responses) did not differ significantly among themselves, as well as being much poorer than the group that had both, highlights the problem of synchronizing the S-R units in a behavioral chain. The advantage in this study of having one appropriate unit without the other is at best negligible. The reason for this is that reinforcement is only achieved consistently when both the appropriate orienting and mediating responses are operating. This particular study points to the need for discovering laws associated with the strengthening and weakening of independent S-R units in a problem solving chain, as well as the principles governing their synchronization.

This study also highlighted a very basic problem in all of these reversal studies. This problem has to do with the very first correct response following the reversal shift. After discovering that the previous mode of responding is erroneous, what makes the subject change his response, i.e., push the button that was previously wrong? Introspective reports fail to provide any clearcut answer and even if they did they would be in need of explaining (Kendler, 1961).

One hypothesis is that the selection of the new correct response is due to the operation of a behavioral chain in addition to the one described in Figure

8. The first nonreinforcement in a reversal shift sets off a chain, the consequence of which is to select the response other than the one that was previously correct. This may result from a number of different reasons (e.g., logical considerations, forgetting, etc.). The important point, however, is that the new key-sorting response occurs contiguously with the implicit mediational response appropriate to a reversal shift. As a result, a new association is formed between the old implicit cue and the new key-pressing response.

In essence, what is being stated is that adult subjects, when making or deciding to make the first correct postshift response, do not adopt the *principle* underlying a reversal shift. Instead, it is assumed processes are operating which encourage the selection of the correct response while an implicit cue appropriate to a reversal shift is operating. This sort of an analysis was described previously (Kendler & Mayzner, 1956) as

> sort of a James-Lange theory of problem solving . . . one makes the overt correct . . . response and if the appropriate symbolic cue is present, then problem solution will occur (pp. 247-248).

Guthrie (1959) says the same thing more neatly: *"What is being noticed becomes a signal for what is being done"* (p. 186).

Again the authors would like to guard against giving the impression of oversimplifying a terribly complex problem. They do not believe the contiguous occurrence of an implicit cue from one chain with the correct overt response from another tells the whole story. This new association in order to persist must be reinforced and in some manner "fit into" the subject's ongoing behavioral chains.

The emphasis on this vertical connection between a cue and a response from different chains is related in a distant way to Hebb's (1958) stressing the role of "chance" in problem solving:

> There are few scientists who have not had the experience of setting out to solve problem A and ending up instead with the solution to B. . . . This is serendipity, the art of finding one thing while looking for another (p. 215).

According to the present analysis, serendipity results from the adventitious and contiguous occurrence of a cue and a response which are themselves segments from different behavior chains. Theoretically it should be possible to demonstrate this point experimentally by training subjects to respond simultaneously to two separate tasks. A problem then would have to be presented that requires for its solution the combination of a stimulus from one chain with the response from the other. In such an experimental situation, controlling the time relationship between the two should have an important effect on problem solving. Presumably contiguity between the two should provide the most optimal conditions for problem solving (Underwood, 1952). The development of this kind of experimental procedure should allow for parametric studies of the basic variables of the phenomenon which has commonly been called "insight," as well as throw light upon issues raised by others (e.g., Cofer, 1957; Maltzman, 1955; Saugstad, 1957).

The pretheoretical model that guides the present research has many more facets that can be exploited. Only one will now be mentioned. Glucksberg (1962), for example, extended neobehavioristic drive theory (Spence, 1956) to problem solving. He used a functional-fixedness problem (Adamson, 1952, Duncker, 1945) in which the correct response in the habit hier-

archy could either be made to be low or high. If the correct habit was low, it would be expected that a strong drive would retard problem solving because it would retard the extinction of the dominant incorrect response (Kendler & Lachman, 1958; Perin, 1942). Since drive energizes behavior, a high drive should facilitate problem solving performance when the correct habit is dominant. The findings were consistent with this analysis.

Because functional-fixedness problems are often represented in perceptual terms, Glucksberg was interested in seeing whether the same drive model could be applied to a simple perceptual recognition problem in which subjects were instructed to identify tachistoscopically presented words as rapidly as possible. The results were similar to those reported for the functional-fixedness study: when the correct response was dominant, an increase in drive improved performance, i.e., the visual duration threshold was lowered. In contrast, increasing drive when the correct response was low in the hierarchy raised the threshold.

There is obviously still much more work, both empirical and theoretical, needed to develop the model that has been described. At this point it may be appropriate to summarize the major points of this paper.

There is not just one way to investigate problem solving. The researcher who is interested in problem solving has several different pretheoretical models from which to choose. This paper reported the results of a research program based on an S-R model in which the importance of horizontal and vertical processes were emphasized. Horizontal processes refer to the linking of successive S-R units into a behavioral chain, while vertical processes refer to the assumption that independent chains occur simultaneously. A series of experiments was reported, the implications of which supported postulating a mediational mechanism within a behavioral chain. By comparing the behavior of human subjects of different ages, as well as relating their results to lower animals, it was possible to infer that as a child matures he makes a transition from responding on the basis of a single unit S-R mechanism to a mediational one. Additional data were cited that suggest the full impact of verbal behavior on problem solving depends on developmental processes that encourage interaction between chains at different vertical levels. It was also suggested that problem solving begins in a simple concept learning task when a correct overt response from one behavioral chain occurs contiguously and adventitiously with the appropriate implicit cue from another chain. The paper was concluded by citing findings that suggested the neobehavioristic drive theory which assumes that the effect of different levels of drive depends on the position of the correct response in the habit hierarchy is applicable to a functional fixedness problem as well as a perceptual-recognition task.

If nothing else, it is hoped that the present paper demonstrates that it is possible to investigate problem solving in a systematic fashion. If more psychologists accepted this possibility and were willing to expend their research energies in the field of problem solving, progress in this area would be greater than it is today.

REFERENCES

ADAMSON, R. E. Functional-fixedness as related to problem solving. *J. exp. Psychol.*, 1952, **44**, 288–291.

BUSS, A. H. Rigidity as a function of reversal and nonreversal shifts in the learning of successive discrimination. *J. exp. Psychol.*, 1953, **45**, 75–81.

BUSS, A. H. Reversal and nonreversal

shifts in concept formation with partial reinforcement eliminated. *J. exp. Psychol.*, 1956, **52**, 162–166.

BUYTENDIJK, F. J. J. Über das Umlernen. *Arch. Néerl. Physiol.*, 1930, **15**, 283–310.

COFER, C. N. Reasoning as an associative process: III. The role of verbal responses in problem solving. *J. gen. Psychol.*, 1957, **57**, 55–58.

DUFORT, R. H., GUTTMAN, N., & KIMBLE, G. A. One trial discrimination reversal in the white rat. *J. comp. physiol. Psychol.*, 1954, **47**, 248–249.

DUNCAN, C. P. Recent research on human problem solving. *Psychol. Bull.*, 1959, **56**, 397–429.

DUNCKER, K. On problem solving. *Psychol. Monog.*, 1945, **58**(5, Whole No. 270).

GLUCKSBERG, S. The influence of strength of drive on functional fixedness and perceptual recognition. *J. exp. Psychol.*, 1962, in press.

GUTHRIE, E. R. *The psychology of learning.* (Rev. ed.) New York: Harper, 1952.

GUTHRIE, E. R. Association by contiguity. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 2. New York: McGraw-Hill, 1959. Pp. 158–195.

HARLOW, H. F. The formation of learning sets. *Psychol. Rev.*, 1949, **56**, 51–65.

HARROW, M., & FRIEDMAN, G. B. Comparing reversal and nonreversal shifts in concept formation with partial reinforcement controlled. *J. exp. Psychol.*, 1958, **55**, 592–597.

HEBB, D. O. *A textbook of psychology.* Philadelphia: Saunders, 1958.

HULL, C. L. Knowledge and purpose as habit mechanisms. *Psychol. Rev.*, 1930, **37**, 511–525.

HULL, C. L. *A behavior system.* New Haven: Yale Univer. Press, 1952.

KELLEHER, R. T. Discrimination learning as a function of reversal and nonreversal shifts. *J. exp. Psychol.*, 1956, **51**, 379–384.

KENDLER, H. H. Problems in problem solving research. In, *Current trends in psychological theory: A bicentennial program.* Pittsburgh: Univer. Pittsburgh Press, 1961.

KENDLER, H. H., & D'AMATO, M. F. A comparison of reversal shifts and nonreversal shifts in human concept formation behavior. *J. exp. Psychol.*, 1955, **49**, 165–174.

KENDLER, H. H., GLUCKSBERG, S., & KESTON, R. Perception and mediation in concept learning. *J. exp. Psychol.*, 1961, **61**, 186–191.

KENDLER, H. H., & KENDLER, T. S. Inferential behavior in preschool children. *J. exp. Psychol.*, 1956, **51**, 311–314.

KENDLER, H. H., KENDLER, T. S., PLISKOFF, S. S., & D'AMATO, M. F. Inferential behavior in children: I. The influence of reinforcement and incentive motivation. *J. exp. Psychol.*, 1958, **55**, 207–212.

KENDLER, H. H., & LACHMAN, R. Habit reversal as a function of schedule of reinforcement and drive strength. *J. exp. Psychol.*, 1958, **55**, 584–591.

KENDLER, H. H., & MAYZNER, M. S., JR. Reversal and nonreversal shifts in card sorting tests with two or four sorting categories. *J. exp. Psychol.*, 1956, **51**, 244–248.

KENDLER, H. H., & VINEBERG, R. The acquisition of compound concepts as a function of previous training. *J. exp. Psychol.*, 1954, **48**, 252–258.

KENDLER, T. S., & KENDLER, H. H. Reversal and nonreversal shifts in kindergarten children. *J. exp. Psychol.*, 1959, **58**, 56–60.

KENDLER, T. S., & KENDLER, H. H. Inferential behavior in children: II. The influence of order of presentation. *J. exp. Psychol.*, 1961, **61**, 442–448.

KENDLER, T. S., KENDLER, H. H., & WELLS, D. Reversal and nonreversal shifts in nursery school children. *J. comp. physiol. Psychol.*, 1960, **53**, 83–88.

KOCH, S. *Psychology: A study of a science.* Vol. 1. *Sensory, perceptual, and physiological formulations.* New York: McGraw-Hill, 1959.

KOHLER, W. *The mentality of apes.* New York: Harcourt, Brace, 1925.

KRECHEVSKY, I. Antagonistic visual discrimination habits in the white rat. *J. comp. Psychol.*, 1932, **14**, 263–277.

LACHMAN, R. The model in theory construction. *Psychol. Rev.*, 1960, **67**, 113–129.

LURIA, A. R. The role of language in the formation of temporary connections. In B. Simon (Ed.), *Psychology in the Soviet Union.* Stanford: Stanford Univer. Press, 1957. Pp. 115–129.

MAIER, N. R. F. Reasoning in humans: I. On direction. *J. comp. Psychol.*, 1930, **10**, 115–143.

MALTZMAN, I. Thinking: From a behavioristic point of view. *Psychol. Rev.*, 1955, **62**, 275–286.

MEYER, M. F. *The fundamental laws of human behavior.* Boston: Gorham, 1911.

PERIN, C. T. Behavior potentiality as a joint function of the amount of training

and the degree of hunger at the time of extinction. *J. exp. Psychol.*, 1942, **30**, 93–113.

SAUGSTAD, P. An analysis of Maier's pendulum problem. *J. exp. Psychol.*, 1957, **54**, 168–179.

SKINNER, B. F. *Science and human behavior.* New York: Macmillan, 1953.

SPENCE, K. W. The nature of discrimination learning in animals. *Psychol. Rev.*, 1936, **43**, 427–449.

SPENCE, K. W. *Behavior theory and conditioning.* New Haven: Yale Univer. Press, 1956.

THOMPSON, R. Successive reversal of a position habit in an invertebrate. *Science*, 1957, **126**, 163–164.

TOULMIN, S. *The philosophy of science.* London: Hutchinson Univer. Library, 1953.

UNDERWOOD, B. J. An orientation for research on thinking. *Psychol. Rev.*, 1952, **59**, 209–220.

WATSON, J. B. Psychology as the behaviorist sees it. *Psychol. Rev.*, 1913, **20**, 158–177.

WODINSKY, J., & BITTERMAN, M. E. Discrimination-reversal in the fish. *Amer. J. Psychol.*, 1957, **70**, 569–575.

(Received June 24, 1960)

Psychological Review
1962, Vol. 69, No. 1, 17–33

A QUANTITATIVE THEORY OF FEELING: 1960 [1]

D. R. KENSHALO AND JOHN P. NAFE

Florida State University

In recent years considerable evidence has been presented which is contrary to the theory of von Frey (1895) that qualitative differences in cutaneous sensations are dependent upon the selective action of specialized, encapsulated end organs. Much of this evidence is from histological studies of skin of various parts of the body and leads to the conclusion that encapsulated end organs cannot be found in hairy skin. Yet hairy skin exhibits sensitivity equal to or greater than that found in glabrous skin (Sinclair, Weddell, & Zander, 1952).

This paper reviews studies concerned with the manner in which nerves terminate in skin and their role in cutaneous sensitivity. The histological studies lead to the conclusion that nerve fibers supplying hairy skin terminate as bare, unencapsulated filaments in relation to three types of tissue: undifferentiated epithelial cells of the dermis and lower layers of the epidermis, hair follicles, and smooth muscle elements of the cutaneous blood vessels. Nafe (1934, 1942) has proposed that these bare filaments are essentially alike and hence respond to a common adequate stimulus, namely, movement either in relation to themselves or to the surrounding tissue. These terminals differ primarily in that they terminate in different tissues of the skin. Both tactile and thermal sensations result when tissue is moved, the former by the direct action of a mechanical stimulus on the tissue, and the latter, by movement of the thermally labile smooth muscle, especially of the arterioles.

Histology of Skin Receptors

Rose and Montcastle (1959) have reaffirmed the position of most textbook writers that the skin senses are mediated by specialized encapsulated endings. Commenting on the recent publications by the Oxford group (Weddell, Palmer, & Pallie, 1955), in which it is maintained that there is but a single type of nerve ending mediating cutaneous sensations, Rose and Montcastle (1959) say:

> If a crisis exists in respect to evaluating the morphology of the endings, it is a crisis of abundance and not of scarcity. One hesitates to accept as a solution to the vexing problem of the morphology of the encapsulated endings a declaration that virtually all morphological differences between them are either insignificant or due to artifacts of the technique (p. 390).

If the Oxford group were the only investigators to arrive at such conclusions, one might expect such a response as quoted above. The controversy as to whether or not encapsulated endings mediate the various qualities of cutaneous sensations has a much longer history.

Three major methods have been employed in the attack on the problem of the histology of cutaneous sensation.

[1] The first paper under this title was published by John P. Nafe in 1929. The present paper was presented in modified form at a symposium on Cutaneous Sensitivity at the Army Medical Research Laboratory, Fort Knox, Kentucky, February 11–13, 1960.

Supported by Contract No. DA 49-007-MD-683 between the Office of The Surgeon General, Department of Army, and Florida State University. Reproduction in part or in whole is permitted for purposes of the United States Government.

TABLE 1

SUMMARY OF HISTOLOGICAL INVESTIGATIONS RELATING SPECIALIZED ENCAPSULATED END ORGANS WITH CUTANEOUS SENSATIONS

Investigator	Skin area		Sensations tested	Encapsulated endings reported
	Glabrous	Hairy		
Type I Investigations				
Donaldson (1885)		forearm	cold	none
Goldscheider (1886)		forearm	cold, warm, touch, pain	none
Häggqvist (1913)		forearm	cold, warm	none
Dallenbach (1927)		upper arm	cold, warm	none
Pendleton (1929)		forearm	cold	none
Woolard (1935)		thigh	cold, pain, touch	none
Bazett (1941)		forearm	cold	none
Weddell (1941)		forearm	cold	Krause end bulb
Belonoschkin (1933)	nipple		cold	complicated plexuses without capsules
Type II Investigations				
von Frey (1895)	conjunctiva		cold	Krause end bulb
	cornea		cold, warm, touch, pain	none
		forearm	cold	Krause end bulb
Strughold (1925)	mouth mucosa		cold	Krause end bulb
Strughold and Karbe (1925)	conjunctiva		cold	Krause end bulb
Bazett et al. (1932)	prepuce		cold, warm, touch	Krause end bulb, Ruffini cylinders
Sinclair et al. (1952)	finger		cold, warm, touch	Meissner corpuscle, Merkel's discs, Krause and Ruffini endings
		forearm	cold, warm, touch	none
		auricle	cold, warm, touch	none
Type III Investigations				
Gilbert (1929)		thigh		none
		epigastrium		none
		chest		none
		breast		none
Gilmer and Haythorn (1941)		back		none
Hagen et al. (1953)		abdomen		none
		finger dorsum		none
Dastur (1955)		232 specimens from hairy skin		none
Winkilmann (1955)		hairy skin		none
Woolard (1936)		thigh		Meissner corpuscle Merkel's disc
Gilbert (1929)	foot sole			Meissner corpuscles
	nipple			Golgi-Mazzoni capsules
Williams et al. (1929)	prepuce			Krause end bulb Ruffini cylinders Meissner corpuscles 4 unnamed types
Cathcart et al. (1948)	nipple			Krause end bulb (rare) unnamed smaller capsules
Gairns (1955), Gairns (1951)	palate			Meissner corpuscles
	gum			Krause end bulb unnamed varieties
Gairns (1953)	tongue			Krause end bulb Meissner corpuscles

The first involves careful sensory testing and permanent marking of the points on the skin which are sensitive to thermal and/or tactile stimulation. After sensory testing and marking, the tissue is removed, stained, and examined microscopically. This is the most direct approach. The second method involves a determination of the number and kind of sensory spots found in a skin area followed by a histological examination of the tissue. The number and type of encapsulated end organs are related to the fre-

quency and kind of sensory spots, but no attempt is made to draw a close relationship between a spot and a particular end organ as in the first method. The third only describes the general location and distribution of various kinds of encapsulated endings. The second and third methods give supportive data at best.

Investigations of the first type have been carried out at least nine times since the first report of Donaldson (1885). With the exception of the study reported by Belonoschkin (1933), these investigations have been carried out on hairy skin. It will be noted in Table 1, with one exception, that the investigators have failed to find encapsulated endings of any sort beneath the spots of stimulation. Investigations of the second type have been conducted primarily on specialized tissue, such as is found in the conjunctiva, the prepuce, and the palmer side of the finger. While there is no direct correlation between a sensory spot and the type of ending beneath it, in general, these investigators conclude that the number and type of encapsulated endings correspond rather well with the number and type of sensory spots reported in that tissue. The list of investigations of the third type included here is by no means exhaustive. The inclusion of these authors is on the basis of their interest in cutaneous sensation. It will be noted that, with one exception, when specimens of skin were removed from the hairy regions of the body, encapsulated endings were not found; whereas, when the specimens of skin were obtained from the nonhairy portions of the body, encapsulated endings were quite prevalent.

In this entire list of investigations there are three exceptions to the general conclusion that encapsulated endings do not occur in hairy skin. These are Weddell (1941), in the first type of investigation; von Frey (1895), in the second type of investigation; and Woolard (1936), in the third type of investigation. Weddell et al. (1955) now contend that the structures which he and Woolard (1936) reported as being encapsulated endings were either unusual sections of hair follicles, or artifacts which resulted from rough handling of the specimens during the fixing process. This appears to be reasonable in view of the results obtained by the other investigators. The conclusion is inescapable that encapsulated endings do not occur in hairy skin. They exist only in specialized skin areas as the conjunctiva of the eye, the mucous membrane of the mouth, the palmer surfaces of the fingers, and the palms of the hand, the soles of the feet, and the genitalia. In hairy skin nerves terminate as bare filaments. The manner of termination of these afferent nerve fibers has been investigated by Weddell, Pallie, and Palmer (1954), and the work of other investigators has been reviewed by Weddell et al. (1955). In general, afferent nerves terminate as bare filaments in relation to three types of structures found in the skin. These are: in and among the cells of the stratum granulosum of the epidermis and among the cells of the dermis; some of the largest fibers serving the skin end in relation to hair follicles, as filaments entwined about the hair shaft; and other smaller fibers terminate in relation to the smooth muscle cells of the cutaneous blood vessels.

Punctate sensitivity of the skin, as described by Blix (1884) and Goldscheider (1884) is often cited as supporting evidence for the specialized encapsulated receptor theory. Some areas or spots of skin are, indeed, more sensitive than others when near threshold intensities of stimulation are used.

However, in studies in which intensity of tactile (Guilford & Lovewell, 1936) or thermal (Jenkins, 1940) stimuli were systematically changed, the number of sensitive spots varied directly with the intensity of the stimuli. It appears that the skin is more or less continuously sensitive to tactile and thermal stimuli, exhibiting peaks and valleys of sensitivity.

CUTANEOUS SENSATION

It is no longer profitable to hypothesize encapsulated end organs which are sensitive to specific types of energy applied to the skin. Another organizing principle must be found to account for qualitative differences in sensations derived from cutaneous stimulation. The filaments in which these nerves terminate appear morphologically alike, although it cannot be maintained that they are identical (Sperry, 1950; Weiss, 1947). On the other hand, there seems to be little reason to assume that the differences are sufficiently great to account for the wide variety of sensory experiences derived from their stimulation. Since there is no reason, at present, to assume that they are different, it would be expected that the terminals of these cutaneous fibers would share a common adequate stimulus, regardless of the tissue in which they end. Evidence has been presented (Nafe & Kenshalo, 1958; Nafe & Wagoner, 1941a, 1941b) which indicates that movement is this common adequate stimulus. Differences in the qualities of cutaneous sensations are based upon differences in the properties of the tissue innervated. Tactile sensations are aroused by and during movement of dermal and epidermal tissue. Thermal sensations are also aroused by tissue movement, although in this instance, the tissue is the smooth muscle composing the walls of the cutaneous vascular system (Nafe, 1942).

The Tactile Sense

Two components exist in the application of any mechanical stimulus to the skin. These are a static component (or magnitude of displacement), and a dynamic component (or velocity and acceleration). Acceleration has not been well investigated, although other work (Hubbard, 1958) seems to indicate that it is of little significance.

The earlier theories concerned with the nature of the adequate stimulus for the tactile sense considered the static component of the mechanical stimulus to be the relevant aspect of that stimulus; thus, there is the pressure theory (Weber, 1846), the gradient theory (Meissner, 1859), and the tension theory (von Frey & Kiesow, 1899). von Frey and Kiesow formulated their theory on the basis of stimuli involving very small areas; that is, stimuli with radii of less than 0.1 millimeter. They found that tension (weight per unit radius) was related to the threshold of sensation for these very small stimuli. On the other hand, for larger stimuli, greater than 0.4-millimeter radius, they were able to relate only the rate of application of the stimulus to the threshold of sensation. Apparently, von Frey and Kiesow were so convinced of the importance of the static component of the mechanical stimulus that they ignored the rate factor involved in the stimuli of larger radii.

Nafe and Wagoner (1941a, 1941b) were the first to record the motion of a mechanical stimulus in producing a tactile sensation. Among the recorded variables of weight, area, time to cessation of the sensation, and the rate of stimulus movement at the cessation of the sensation, a relationship was es-

tablished between weight and the rate of stimulator movement at the time of cessation of the tactile sensation. The rate of movement varied independently of the area at the time of cessation of sensation; hence, any aspect of the stimulus which involves this variable, e.g., pressure or tension, must be considered to be of no importance in producing and maintaining a tactile sensation. In addition, they found that depth of deformation of the tissue was a function of the applied pressure and must also be rejected as the stimulus. In other words, they were not able to relate any of the static aspects of the stimulus to the point at which cessation of the sensation occurred. As a result, they concluded that the dynamic component was the relevant aspect of the stimulus; further, that cessation of the tactile sensation was the result of failure of the dynamic component of the mechanical event to stimulate, rather than failure of the receptor to respond to the stimulus. The rate of tissue deformation had reached a subliminal value.

To test this notion, Nafe and Wagoner (1941b) lowered a weight of 17.5 grams onto the skin is 8.75-gram increments. The subject reported a tactile sensation with the application of each increment and cessation of the sensation when the rate of stimulus movement became small. Furthermore, when the second 8.75-gram weight was removed, the subject again felt a tactile sensation which disappeared at approximately the same rate of tissue movement as before. They contend that had the receptors, rather than stimulation, failed, the subject would not have felt the application of the second weight, nor would he have felt its removal.

In order to locate this effect in the peripheral receptor mechanism, rather than in the central nervous system,

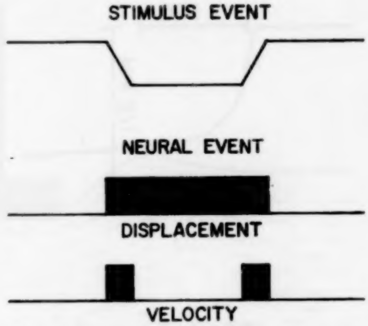

FIG. 1. Hypothetical response in a nerve fiber to a stimulus. (If the tactile receptor responds to the static component of a stimulus, action potentials should occur throughout stimulation as shown in the second line labeled "Displacement." The neural reponse to the dynamic component of the stimulus, in the third line, will last only during movement of the stimulus.)

Nafe and Kenshalo (1958) repeated Nafe and Wagoner's experiments using a slightly modified version of the Nafe and Wagoner apparatus. In place of the human observer, an animal, usually a rat, was used. Instead of a verbal report by the subject, action potentials were recorded from the nerve innervating the stimulated tissue. Simultaneous recordings of the stimulus action and the neural discharge were made on photographic paper. In most of the preparations, the tongue of the rat was stimulated and recordings were made from the lingual branch of the fifth cranial nerve. Some preparations were made of the shaved leg of the rat and records were made from the femoral nerve, while in other experiments, the skin of frog was used.

Figure 1 shows the type of record which would be predicted if the receptors of the skin respond to either the static component (displacement) or the dynamic component (velocity) of the mechanical stimulus.

Figure 2 shows a record of the stim-

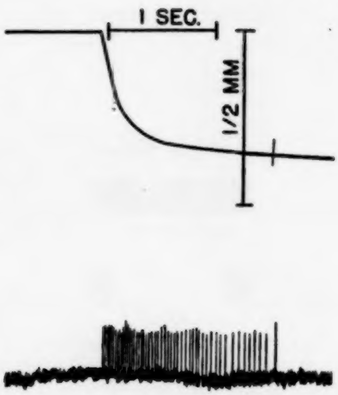

Fig. 2. Action potentials in the lingual nerve in response to weight applied to the tongue of rat. (Movement of the weight into the tissue is shown in the top line. This is comparable to only the left half of Figure 1. In other experiments the weight was later lifted. Action potentials were also obtained during its removal—modified from Nafe & Kenshalo, 1958.)

ulus resulting from an 8.75-gram weight sinking into the tissue of rat tongue, and the neural discharge which results. In this type of recording, the stimulator follows the adjustment of the tissue to the weight rather closely. It can be seen that the neural discharge which accompanied application of the stimulator stopped before the tissue had completely adjusted to the added weight. The stimulator was still moving into the tissue, in this instance at the rate of 0.019 mm/sec, when the neural discharge stopped. Records such as these establish the point that the cessation of sensation associated with a minimal rate of movement is a peripheral event.

The question still remains as to whether the cessation of discharge is a function of the stimulus effectiveness alone, or if it is in part due to adaptation by the receptor. To test this, the stimulus was hydraulically driven into the tissue in two steps of equal velocity, but the second step was of less magnitude than the first. If the cessation of discharge were associated with receptor adaptation rather than a stimulus parameter, the neural discharge associated with the second step should be less than that of the first. On the other hand, if rate of movement is the determining condition, the discharge to the second step should be equal to

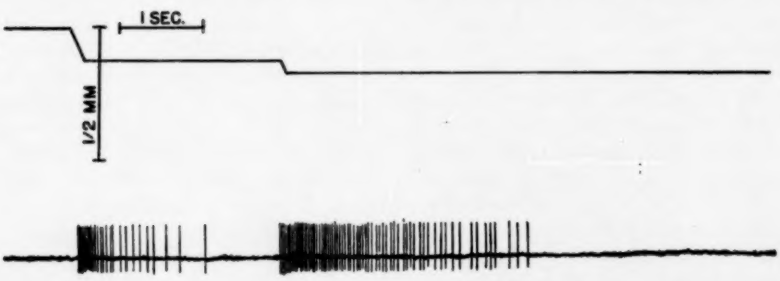

Fig. 3. Action potentials in the lingual nerve of rat when the stimulator is driven hydraulically into the tongue tissue in two steps of equal velocity. (The course of the stimulus into the tissue is shown in the top line—modified from Nafe & Kenshalo, 1958.)

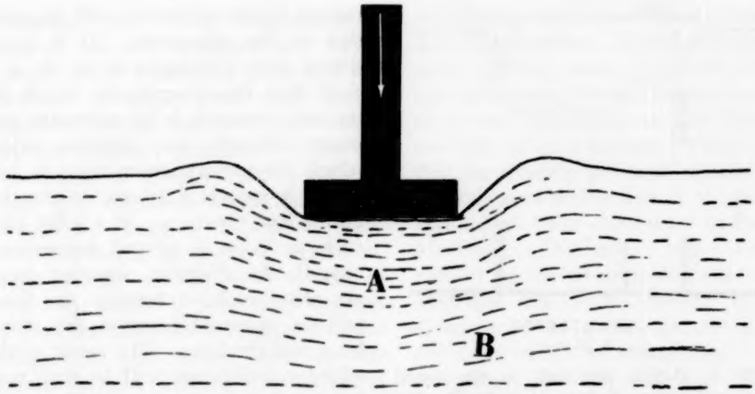

FIG. 4. Hypothetical cross section through skin showing mechanoreceptors at A and B in relation to the stimulator.

or greater than that from the first. As can be seen from Figure 3, the second hypothesis is confirmed. The longer discharge associated with the second step is explained by greater mechanical stress in the tissue following the application of the first step, than prior to its application. Thus, in Figure 4, if it is assumed that the terminals of the fiber from which the record was made are at A, it is apparent that movement of tissue at A will be greater per unit movement of the stimulator when A is under mechanical stress than when it is not. The observations of Loewenstein (1956) emphasize the importance of the mechanical factors involved in the stimulation of skin receptors. He has reported that the duration of a single unit discharge in response to a standard stimulus in frog skin is greater when the skin is placed under mechanical stress than when it is flaccid. Our own observations of a decrease in the velocity of stimulus movement with increasing weights applied over a constant area, may be explained in a similar fashion. One might think of the relation of the stimulus and receptor as analogous to a mechanical gear train with a large amount of backlash. As the gears are put under mechanical stress, by loading, e.g., the use of spring loaded antibacklash gears, the output of the gear train follows that of the input more closely. Similarly the tissue at A follows movements of the stimulator more closely.

To test these notions it seemed appropriate to use an unyielding tissue. The coupling between the stimulator and receptor would be closer and tissue readjustments to the stimulus action would be minimal compared to the rat's tongue. Frog skin is suited for such an investigation. As can be seen in Figure 5, the neural discharge occurred only during the time that the tissue was being moved. The second interesting aspect is that when the stimulator was removed from the frog's back a third series of neural impulses occurred.

Movement of tissue appears to be the adequate stimulus for nerve terminals found in a variety of locations. Nafe and Wagoner (1941b) used the skin of the web between the thumb and index finger and just above the knee. Nafe and Kenshalo (1958) used the tongue and hind leg of rat,

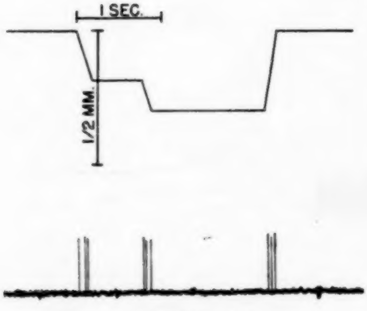

FIG. 5. Action potentials in the dorsal cutaneous nerve of frog as a result of a stimulus driven hydraulically into the skin and then removed. (The course of the stimulus into the tissue is shown in the top line—modified from Nafe & Kenshalo, 1958.)

and frog skin. There appears to be little difference in the mechanism of stimulation of the nerve terminals which exist in these various locations. Differences which have been observed may be accounted for by differences in the mechanical properties of the tissue surrounding them. It has been known for some time that nerve terminals surrounding hair follicles respond only to movement of the hair (Adrian, 1930). In addition, the works of Gray and Malcolm (1950), Gray and Matthews (1951), and Gray and Sato (1953) have demonstrated that the Pacinian corpuscle is also responsive to only the dynamic aspect of the stimulus—movement.

According to this concept, tactile adaptation, in the sense of a protracted but diminishing sensation or nerve discharge resulting from a constant stimulus, becomes meaningless. Consider the diagram of Figure 4. The rate of tissue movement or adjustment at A will be much less than the rate of movement of the stimulator, but as it continues to sink into the tissue, the rate of tissue movement will approach that of the stimulator. If it is assumed that a receptor is at A, it is clear that the stimulus to which the receptor responds is by no means constant, although the stimulus object, which the experimenter sees, is constant—it weighs 5 grams, or exerts 5 grams/cm^2 pressure. If a solid plate such as bone is placed immediately beneath the diagram, shearing forces are also produced because the tissue will be squeezed between the stimulator and the bone. The result of this added complication will be that resolution of these stresses will require much longer at A than at B. According to the older electrophysiological literature, in comparing the neural discharges from receptors at A and B, it would have been said that A was slow adapting while B was fast adapting, implying a difference in the neurological functioning of each rather than in tissue mechanics. An important parameter of the problem is overlooked if the mechanical properties of the tissue are not also investigated.

The Thermal Sense

Before considering proposals for thermal receptors, the facts concerning thermal sensation should be examined. They impose rather stringent requirements upon any theory concerned with the nature of thermal reception. The thermal sense is different from other senses. In the other senses absence of sensation is associated with near zero energy levels. Zero thermal energy is at absolute zero, approximately $-273°$ C; yet, psychological zero or zero thermal sensation is at skin temperature—approximately 33° C.

Stimulation of thermal receptors depends upon a temporal thermal gradient. Within limits, thermal sensations occur only when the stimulus temperature change occurs at above a

minimal rate. Skin temperature is known to shift spontaneously by as much as 0.5° within the space of a few minutes (DuBois, 1941); yet during these shifts the subject is not aware of any thermal sensation. The observations of Hensel (1950) indicate that skin temperature may be shifted within the range of 29°–36° C without producing a thermal sensation, provided the change is at a rate of less than 0.007° C/sec. This appears to be the zone of complete thermal adaptation and hence specifies the limits of the shifting psychological zero. The shifting zero point accounts for the frequent observation that the thermal reception system in man is not a good absolute thermometer. While thermal sensations do not occur with large changes in skin temperature at a slow rate, threshold sensations occur at a rate of 0.001° C/sec for 3 seconds (Hardy & Oppel, 1936).

The thermal threshold is affected by the temperature of the skin (Hensel, 1950). Ebaugh and Thauer (1950) have reported that the cool threshold is increased at skin temperatures above 33° C. Kenshalo, Nafe, and Dawson (1960) confirmed their findings and, in addition, were able to show that the warm threshold decreases at skin temperatures above 34° C.

Theories which depend on a spatial thermal gradient must be rejected. Ebbecke (1917) and Bazett and McGlone (1932a) have suggested spatial gradient theories. The notions of Bernhard and Granit (1946) and von Euler (1947), that nerve itself may act as a thermal receptor, is also based on regional cooling of the nerve. Lele, Weddell, and Williams (1954) have extended this notion and suggested a Bolometer Theory which depends on warming and cooling of terminals in relation to their stem axons. Rejection of spatial gradients as a means of stimulating thermal receptors requires some justification. Hensel and Zotterman (1951c), using electrophysiological techniques, located the termination of a fiber responding to cold close to one side of cat's tongue. They found that the receptor performs equally well regardless of which side of the tongue was cooled. The older work of Bazett, McGlone, Williams, and Lufkin (1932) on the prepuce, likewise, demonstrates the ineffectiveness of the direction of the spatial thermal gradient as a means of stimulating warm receptors. Vendrik and Vos (1958) abolished a spatial thermal gradient altogether by warming the skin with a radar beam of 10-centimeter wave length, yet were unable to observe any appreciable difference in the warmth threshold from that obtained by infrared radiation. Electromagnetic wave lengths of this order are quite penetrating, according to them, and warm the deeper tissues as rapidly as the skin surface. If a spatial thermal gradient were necessary to stimulate the thermal sensing mechanism, they should have obtained a much higher thermal threshold using the radar beam than by infrared stimulation. The latter raises the skin surface temperature and depends upon conduction of the thermal energy to the deeper tissues.

These experiments demonstrate the ineffectiveness of spatial thermal gradients in depth as a stimulus for the thermal sensing mechanism. It is still possible that gradients across the surface of the skin may be of some importance, although it seems unlikely. Experiments such as those by Hardy and Oppel (1936) in which large areas of skin, e.g., the torso or torso plus face, were warmed uniformly by radiant energy should have produced thermal sensations only at the edges of the stimulus. Their subjects made

no such observation. Similarly, the thermal sensation aroused by laying a cold coin on the skin should occur only at the margin, but this is not consistent with common observation.

Two types of receptor mechanisms have been proposed for mediating thermal sensations. One assumes that cutaneous nerve terminals are differentially sensitive, some responding only to mechanical energy, while others respond to thermal energy. The other maintains that endings of the skin respond to a common adequate stimulus—movement. Thermal reception would then be by means of tissue movement resulting from thermal stimulation. It is well known that smooth muscle is thermally responsive, contracting when cooled and relaxing when warmed. Such movement induced in smooth muscle of the cutaneous vascular system would produce impulses in the nerve fibers terminating among these smooth muscle elements.

The notion that nerve terminals respond differentially to mechanical and thermal energy is supported by evidence from Hensel and Zotterman (1951b). They found that stem axons which show activity in response to the application of mechanical stimuli to their terminals do not respond well to thermal stimuli and vice versa. Their findings indicate that mechanically sensitive nerve terminals have stem axons of larger diameter (larger than 8 micra); whereas, those responding to thermal stimuli are smaller (1.5–5 micra). They interpret their findings as indicating that nerve terminals are inherently different in their responsiveness to thermal and mechanical energy. However, such differences could occur as a result of differences in the mechanical and thermal properties of the tissue in which the nerve terminals end.

Several mechanisms have been suggested to account for the apparent differential sensitivity of nerve terminals. Bishop (1946) has related it to characteristics associated with the size of the stem fiber involved. Others have suggested that the nerve terminals differ in their sensitivity to chemical changes. Bazett and McGlone (1932b) proposed that thermal sensations might be the result of alterations in blood acidity produced by warming and cooling. Consistent with this, Dodt (1956) has demonstrated a differential effect of CO_2 tension on the response of nerves stimulated thermally. In his Concentration Theory, Jenkins (1941) has also suggested a chemical intermediary. Bare nerve endings, according to Jenkins, would discriminate between a catabolic and an anabolic phase of some chemical reaction brought about by a change in temperature. Presumably bare nerve endings responding to mechanical stimuli would be insensitive to these changes.

The suggestion of Hensel and Zotterman (1951b) on the mechanism of the stimulation of cold receptors is reminiscent of the anabolic-catabolic hypothesis. Because of the rather complex shifts in the frequencies of impulses which they found in fibers responding to a temperature reduction, they suggest that there are two temperature dependent processes, probably chemical, involved in the stimulation of cold fibers (Zotterman, 1959). The impulse frequency resulting from cold stimulation is a function of a difference in the activity of these two processes. Although they do not speculate on the neural response to warming, one must presume at least one, and probably two, additional processes are necessary. Dodt (1953) has shown that the responses of mammalian A fibers to warming and cooling are

functions of their diameter. The differences do not appear great enough, however, to aid in understanding Hensel and Zotterman's suggestion of a chemical process controlling the responsiveness of certain fibers to thermal stimuli.

Hypotheses which attempt to explain differences between thermal and mechanical sensitivity on the basis of nerve terminals of different sensitivity must assume two different sets of cutaneous nerve terminals, one set responding to a mechanical form of energy, the other to thermal energy. If the difference between the two sets of terminals is a chemical one, the chemical reaction should be one which is in equilibrium at the mid-range of the thermal energy continuum, which may have its thermal equilibrium point shifted by 6 or 7 degrees if done sufficiently slowly, yet which will become unstable when changed rapidly by as little as 0.003° C (Hardy & Oppel, 1936). To further complicate the theory, Hensel and Zotterman (1951b) find that of the terminals responding to thermal energy, some are activated only by an elevation, while others respond to a reduction in the thermal energy. Thermal fibers must now be endowed with even greater selectivity, either to discriminate between two different chemical reactions, one for warming and one for cooling, or between the directions in which a reversible chemical reaction is driven. In addition, the response of terminals to cooling is to be explained by the interacting of two temperature dependent processes, according to Zotterman (1959).

An alternative, involving fewer assumptions, is that nerve fibers and their terminals are largely alike; further that the differentiation necessary to account for thermal and mechanical responsiveness occurs as a direct result of the tissue in which the fibers terminate. It has been demonstrated that nerve terminals, similar to those terminating among the cells of the epidermis, end among the smooth muscle fibers of the cutaneous vascular network (Weddel et al., 1954; Woolard, 1926). Smooth muscle is known to be thermally responsive, relaxing upon warming and constricting when cooled. Furthermore, Nafe and Kenshalo (1958) have demonstrated that movement of the nerve terminal either upon itself or in relation to its surrounding tissue best fits the facts concerning mechanical reception.

Goldscheider (1886) was probably the first to suggest a relationship between the cutaneous vascular system and thermal sensation. He noted in his biopsies of temperature sensitive spots the "striking and immediate proximity" of blood vessels to cold spots which he has previously mapped. Elaborating on this, Nafe (1934) pointed out the striking relationship between changes in the smooth muscle response to temperature and the subjective reports of stimulation by the same temperatures. The figures which Nafe (1934) cited applied to smooth muscle of the uterus and while they apply to smooth muscle in general, recent evidence indicates that the smooth muscle of the cutaneous vascular network, especially in the venules,

> show an extraordinarily prominent vasomotor activity in contrast to other tissues, such as mesenteric structures, the wall of the intestinal tract, and skeletal muscle, where venous vasomotion is minimal and seemingly passive (Zweifach, 1959, p. 69).

Furthermore, the venular tributaries in the skin are ten- to twenty-fold more sensitive to a fall in temperature of 1–2° C than other tissues (Zweifach, 1959). While the Vascular Theory is usually thought of as involving only the arterioles the responsiveness of

venular smooth muscle and that of the hair erector muscles should not be overlooked.

The Vascular Theory accounts for the requirements of a thermal theory as outlined at the beginning of this section. It does not depend upon spatial thermal gradients to achieve stimulation and is uniquely suited for explanations involving temporal thermal gradients. Reference was made earlier in the paper to the fact that in order to stimulate nerve terminals located among the cells of the epidermis and hair follicles, minimal rates of tissue movement were necessary. On the assumption that nerve terminals are essentially alike, the same conditions of stimulation should apply to the nerve terminals ending in relation to the fibers of the smooth muscle of the cutaneous vascular system. The shifting psychological zero with no attendant thermal sensation is accounted for in terms of the constriction or dilation of the cutaneous vessels occurring slowly enough to be below the threshold rate of movement necessary to excite the nerve terminals in the vessel walls. Complete thermal adaptation represents the adjustment of vascular relaxation or constriction from one level to another; and the range through which Hensel (1950) has reported that complete thermal adaptation is possible, represents the range through which the smooth muscle of the cutaneous vessels can adjust and adopt a relatively stable end point. Outside of this range vascular smooth muscle would be unable to achieve a relatively stable adjustment to the temperature.

Changes in the size of the thermal threshold as a function of the skin temperature should be expected since the latter would govern the degree of relaxation or constriction of the smooth muscle. Predictions of the direction in which the thresholds would go are difficult to make because of ignorance of the mechanical and physiological characteristics of afferent innervation of smooth muscle.

In addition to fitting these requirements, the theory also handles facts which have been somewhat perplexing to other theories. Thermal sensations occur under some circumstances which do not include a change in the environmental temperature. It has been noted, especially in the earlier work on the thermal sense, that mechanical stimulation of the skin may produce sensations of warm or cool (Nafe, 1934). Smooth muscle is known, upon occasion, to relax or contract when prodded, depending upon its state of contraction at the moment (Evans, 1956). Thermal sensations resulting from mechanical stimulation can be explained in this manner if one assumes that they are mediated by smooth muscle of the cutaneous vascular system. Warm sensations accompanying a blush occur as a result of chemically mediated vasodilation since there is apparently little if any reflex vasomotor control in the "blush area" (Burton, 1959). The chill accompanying scratching a fingernail across a blackboard is the result of a reflexly aroused general vasoconstriction.

It is not necessary to postulate different types of free nerve endings, each responsive to one type of energy; nor is it necessary to conceive of two different systems, one for warm and one for cool. Two possible explanations exist. One depends upon neural patterns while the other depends upon different fibers being stimulated, based upon the way in which they terminate among the smooth muscle elements. In the first instance neural patterns aroused by constriction could conceivably differ from those aroused by

relaxation. Patterns in other systems result in different sensations. For example, the peristaltic waves in the stomach passing from the cardiac to the pyloric valves are associated with hunger pangs, while those same contractions passing in the opposite direction result in the entirely different sensation of nausea. The direction of travel is simply a matter of timing. In addition to central factors involved in thirst, the patterns aroused in the membranes of the throat when they are dry, in comparison to when they are wet, contribute markedly to this sensation. The urge to micturate and to defecate are likewise results of patterning of neural discharges.

On the other hand, Hensel and Zotterman (1951c), using electrophysiological methods on the cat's tongue, have described two apparently different sets of nerve fibers, one responding to warm and the other to cool stimuli. In striated muscle, such systems occur; one responding when the muscle is contracted and the other when it is stretched. In the first instance the receptor elements are connected in series with the muscle fibers; whereas, in the latter case they are connected in parallel with the muscle fiber. Whether the same type of endings exist in smooth muscle has not been investigated, although Fischer (1944) has described some afferent terminals in smooth muscle as "spray-like," and others as "spindle-like."

Hensel and Zotterman (1951b) have described fibers responding to thermal stimulation as smaller (1.5–5 micra) than those responding to tactile stimulation (8 micra and larger). Weddell and Pallie (1954) have shown that many of the fibers innervating the cutaneous blood vessels are of dorsal root origin. In addition, these fibers are among the smallest found in the skin, being, for the most part, less than 3 micra in diameter. The fibers ending in cutaneous vessels, then, correspond with the fiber spectrum which Hensel and Zotterman have described as responsible for thermal sensitivity.

One of the most frequently cited reasons for considering warm and cool to be mediated by different receptor mechanisms is the long reaction time of an observer to warm compared with cool stimuli. Bazett et al. (1932) and Hensel and Zotterman (1951a) have used this difference in latency to calculate the depth of the receptors. These calculations are based on the assumption that the latency of response of the receptors are the same but occur at different depths. An alternate explanation, in terms of the Vascular Theory, is equally possible. Constriction is an active process requiring little time to accomplish. On the other hand, dilation is a passive response and requires a much longer time to reach a threshold rate of movement.

Paradoxical cold provides little difficulty for the Vascular Theory. Paradoxical cold is a term ascribed to the situation in which a cold sensitive area is stimulated by means of a high temperature and produces a cold sensation. At high temperatures smooth muscle has constricting elements within a generally relaxing system (Evans, 1956). It might be expected that an area normally responding to a cool stimulus would also produce a cool sensation when stimulated by a high temperature. It is interesting to note that Zotterman (1953) reports fibers which normally respond to cooling, after a period of no response to temperature above 35° C. suddenly start to respond when stimulated by temperatures of 45° C. and higher.

Sensations of heat are the result of a mixed sensation of warm and paradoxical cold according to Alrutz

(1900). It is known (Evans, 1956) that when smooth muscle is warmed to above approximately 45° C. elements of the muscle begin to constrict while the main body of the muscle continues to relax. According to the Vascular Theory hot sensations result from the neural activity brought about by the constricting elements in the generally relaxing smooth muscle of the cutaneous arterioles.

The question arises as to how the sensorium distinguishes between tissue movements induced by mechanical stimuli and those induced by thermal stimulation. Evidence has already been presented to indicate that there is strong likelihood that different afferent nerve fibers respond to mechanical stimulation than to thermal. Terminals ending in the cutaneous vascular system, because of their smaller size and probably also because of the mechanical properties of the tissue of blood vessels, do not respond as readily to mechanical stimuli, as those which terminate among the cells of the epidermis, dermis, and in relation to hair follicles. There is no contention that nerve or its terminals may not be stimulated directly by a temperature change, but it has never been demonstrated that their temperature sensitivity is sufficient to account for thermal sensations in the skin. The coding system is based on different fibers being activated by different types of energy as the result of the mechanical and thermal properties of the tissue in which their terminals end. Further differentiation is accomplished in the spinal cord. Descriptions of the neuroanatomy of the spinal cord (Ranson & Clark, 1959) maintain that fibers serving touch travel in the fasciculi gracilis and cuneatus as well as in the ventral spinothalamic tract, while those allied with thermal sensation and pain ascend the cord predominantly in the lateral spinothalamic tract, although there is probably considerable intermingling of fibers among the various tracts (Rose & Montcastle, 1959). Fibers from these tracts terminate in different brain stem nuclei but their connections are beyond the scope of this paper.

Aside from the fact that the theory can account for a large number of the facts concerning tactile and thermal sensitivity it has the added advantages of simplicity and parsimony. It can serve as a convenient model within which to organize additional facts as well as bringing into sharp relief gaps in present knowledge. Furthermore, such a model allows predictions and suggests direct experiments which should be done.

SUMMARY

Qualitative theories requiring specialized encapsulated end organs or specialized peripheral nerve fibers to account for the various qualities (pressure, warm, and cold) of cutaneous sensation fail to present direct evidence that such specialization exists in peripheral nerves or in their terminations. The Quantitative Theory, discussed here, maintains that qualities of cutaneous sensation are, in part, a function of the mechanical and thermal properties of the tissue in which the cutaneous afferent nerves terminate and variations in the temperal and spatial patterns of the afferent nerve discharge. The latter variations are manifest in the frequency and duration of the impulses, and the area over which they arise as well as the relative number of fibers actuated within the area. The Quantitative Theory consists of four postulates.

1. Nerves which end freely, without encapsulation of any sort, are the primary neurological elements present in hairy skin capable of mediating

sensations derived from normal stimulation of the skin. Nerves ending freely in the skin terminate in relation to different tissues.

a. Some fibers branch and terminate among the dermal and epidermal cells of the stratum granulosum.

b. Others, and probably some branches of those terminating as in (a) form a complex net about the hair follicles.

c. Smaller fibers branch and terminate among the smooth muscle elements in the walls of the cutaneous vascular system and perhaps the hair erector muscles.

2. The terminals of nerve fibers supplying the skin are functionally similar, differing primarily by the tissue in which they terminate. Being so, they share a common adequate stimulus which is movement of the terminals either upon themselves or in relation to their immediately surrounding tissue.

3. Smooth muscle responds to thermal changes, in general, relaxing when warmed but constricting when cooled.

4. The thermal and mechanical properties of the tissue in which the terminals end determine which fibers will respond to mechanical stimulation and which will be thermally sensitive. Distinction between thermal and tactile sensations depends upon the central connections which the fibers make. Qualitative differences in sensations derived from various mechanical stimuli are the result of differing patterns of neural excitation. Thus stroking a smooth surface results in different movement patterns in the tissue and hence different patterns of action potentials from stroking a rough surface. Distinction of warm and cool sensations results from either different patterns of action potentials caused by vasoconstriction as compared to vasodilation, or different fiber terminals responding to vasoconstriction than to vasodilation.

REFERENCES

ADRIAN, E. D. The effects of injury on mammalian nerve fibers. *Proc. Roy. Soc. Lond., Ser. B,* 1930, **106**, 596–618.

ALRUTZ, S. Studien auf dem Gebiete den Temperatursinne. *Skand. Arch. Physiol.,* 1900, **10**, 340.

BAZETT, H. C. Temperature sense in man. In, *Temperature: Its measurement and control in science and industry.* New York: Reinhold, 1941. Pp. 489–501.

BAZETT, H. C., & MCGLONE, B. Studies in sensation: II. The mode of stimulation of cutaneous sensations of cold and warmth. *Arch. Neurol. Psychiat.,* 1932, **27**, 1031–1069. (a)

BAZETT, H., & MCGLONE, B. Studies in sensation: III. Chemical factor in the stimulation of end-organ giving temperature sensations. *Arch. Neurol. Psychiat.,* 1932, **28**, 71–91. (b)

BAZETT, H. C., MCGLONE, B., WILLIAMS, R. G., & LUFKIN, H. M. Sensation: I. Depth, distribution and probable identification in the prepuce of sensory end-organs concerned in sensations of temperature and touch; thermometric conductivity. *Arch. Neurol. Psychiat.,* 1932, **27**, 489–517.

BELONOSCHKIN, B. Ueber die Kaltreceptoren der Haut. *Z. Biol.,* 1933, **93**, 487–489.

BERNHARD, C. G., & GRANIT, R. Nerve as model temperature end-organ. *J. gen. Physiol.,* 1946, **29**, 257–265.

BISHOP, G. Neural mechanisms of cutaneous sense. *Physiol. Rev.,* 1946, **26**, 77–102.

BLIX, M. Experimentelle Beiträge zur Losung der Frage über die specifische Energie der Hautnerven. *Z. Biol.,* 1884, **20**, 141–156.

BURTON, A. C. Physiology of cutaneous circulation, thermoregulatory functions. In S. Rothman (Ed.), *The human integument.* Washington, D. C.: American Association for the Advancement Science, 1959.

CATHCART, E. P., GAIRNS, F. W., & GARVEN, H. S. D. The innervation of the human quiescent nipple, with notes on pigmentation, erection, and hyperneury. *Trans. Roy Soc. Edinb.,* 1948, **61**, 699.

DALLENBACH, K. The temperature spots and end-organs. *Amer. J. Psychol.,* 1927, **39**, 402–427.

DASTUR, D. K. Cutaneous nerves in leprosy: The relationship between histopathology and cutaneous sensibility. *Brain*, 1955, **78**, 615–633.

DODT, E. Differential thermosensitivity of mammalian A-fibres. *Acta physiol. Scand.*, 1953, **29**, 91–108.

DODT, E. Die Aktivität der Thermorecep-toren bie nicht-thermischen Reizen bekannter thermoregulatorischer Wirkung. *Pflüg. Arch. ges. Physiol.*, 1956, **263**, 188–200.

DONALDSON, H. H. Research on the temperature sense. *Mind*, 1885, **10**, 399–416.

DUBOIS, E. F. The temperature of the human body in health and disease. In, *Temperature: Its measurement and control in science and industry.* New York: Reinhold, 1941.

EBAUGH, F., & THAUER, R. Influence of various environmental temperatures on the cold and warmth thresholds. *J. appl. Physiol.*, 1950, **3**, 173–182.

EBBECKE, U. Über die Temperaturemfindungen in ihrer Abhängigkeit von der Hautdurchbluting und von den Reflexzentren. *Pflüg. Arch. ges. Physiol.*, 1917, **169**, 395–462.

EVANS, C. L. *Principles of human physiology.* (12th ed.) Philadelphia: Lea & Febiger, 1956.

FISCHER, E. Vertebrate smooth muscle. *Physiol. Rev.*, 1944, **24**, 467–490.

GAIRNS, F. W. The sensory nerve endings of the human palate. *Quart. J. exp. Physiol.*, 1955, **40**, 40–48.

GAIRNS, S. W. Nerve endings in the human gum and hard palate. *J. Physiol.*, 1951, **115**, 70.

GAIRNS, S. W. Sensory nerve endings other than taste buds in the human tongue. *J. Physiol.*, 1953, **121**, 33.

GILBERT, R. W. Dermal sensitivity and the differentiated nerve terminations of the human skin. *J. gen. Psychol.*, 1929, **2**, 445–459.

GILMER, B. VON H., & HAYTHORN, S. R. Cutaneous pressure-vibration spots and their underlying tissues. *Arch. Neurol. Psychiat.*, 1941, **46**, 621–628.

GOLDSCHEIDER, A. Die specifische Energie der Temperaturnerve. *Mschr. prakt. Dermatol.*, 1884, **3**, 53–76.

GOLDSCHEIDER, A. Histologische Untersuchungen über die Endigungen der Hautsinnesnerven beim Menschen. *Arch. Anat. Physiol., Physiol. Section*, 1886, 191–227.

GRAY, J. A. B., & MALCOLM, J. L. Initiation of nerve impulses by mesenteric pacinian corpuscles. *Proc. Roy. Soc. Lond., Ser. B*, 1950, **137**, 96–114.

GRAY, J. A. B., & MATTHEWS, P. B. C. Response of pacinian corpuscles in the cat's toe. *J. Physiol.*, 1951, **113**, 475–482.

GRAY, J. A. B., & SATO, M. Properties of the receptor potential in pacinian corpuscles. *J. Physiol.*, 1953, **122**, 610–636.

GUILFORD, J. P., & LOVEWELL, E. M. The touch spots and the intensity of the stimulus. *J. gen. Psychol.*, 1936, **15**, 149–159.

HAGEN, E., KNOCHE, H., SINCLAIR, D., & WEDDELL, G. The role of specialized nerve terminals in cutaneous sensibility. *Proc. Roy. Soc. Lond., Ser. B*, 1953, **141**, 279–287.

HÄGGQVIST, G. Histophysiologische Studien über die Temperatursinne der Haut des Menschen. *Anat. Anz.*, 1913, **45**, 46–63.

HARDY, J., & OPPEL, T. The thermal response of skin to radiation. *Physics*, 1936, **7**, 466–479.

HENSEL, H. Temperaturempfindung und intracutane Wärmebewegung. *Pflüg. Arch. ges. Physiol.*, 1950, **252**, 165.

HENSEL, H., & ZOTTERMAN, Y. Action potentials of cold fibres and intracutaneous temperature gradient. *J. Neurophysiol.*, 1951, **14**, 377–385. (a)

HENSEL, H., & ZOTTERMAN, Y. Quantitative Beziehungen zwischen der Entladung einzelner Kaltfasern der Temperatur. *Acta physiol. Scand.*, 1951, **23**, 291. (b)

HENSEL, H., & ZOTTERMAN, Y. Response of mechanoreceptors to thermal stimulation. *J. Physiol.*, 1951, **115**, 16–24. (c)

HUBBARD, S. J. A study of rapid mechanical events in a mechanoreceptor. *J. Physiol.*, 1958, **141**, 198–218.

JENKINS, W. Studies in thermal sensitivity: XIV. Further evidence on the effects of stimulus temperature. *J. exp. Psychol.*, 1940, **27**, 76.

JENKINS, W. Studies in thermal sensitivity: XVII. The topographical and functional relations of warm and cold. *J. exp. Psychol.*, 1941, **29**, 511–516.

KENSHALO, D. R., NAFE, J. P., & DAWSON, W. W. A new method for the investigation of thermal sensitivity. *J. Psychol.*, 1960, **49**, 29–41.

LELE, P. P., WEDDELL, G., & WILLIAMS, C. M. The relationship between heat transfer, skin temperature and cutaneous sensibility. *J. Physiol.*, 1954, **126**, 206–234.

LOEWENSTEIN, W. R. Modulation of cutaneous mechanoreceptors by sympathetic stimulation. *J. Physiol.*, 1956, **132**, 40–60.

MEISSNER, G. Untersuchunger über den Tastsinn. *Z. rat. Med.*, 1859, **7**, 92–118.

NAFE, J. P. A quantitative theory of feeling. *J. gen. Psychol.*, 1929, **2**, 199–211.

NAFE, J. P. The pressure, pain, and temperature senses. In C. A. Murchison (Ed.), *A handbook of general experimental psychology.* Worcester, Mass.: Clark Univer. Press, 1934. Pp. 1037–1087.

NAFE, J. P. Toward the quantification of psychology. *Psychol. Rev.*, 1942, **49**, 1–18.

NAFE, J. P., & KENSHALO, D. R. Stimulation and neural response. *Amer. J. Psychol.*, 1958, **71**, 199–208.

NAFE, J. P., & WAGONER, K. S. The nature of pressure adaptation. *J. gen. Psychol.*, 1941, **25**, 323–351. (a)

NAFE, J. P., & WAGONER, K. S. The nature of sensory adaptation. *J. gen. Psychol.*, 1941, **25**, 295–321. (b)

PENDLETON, C. R. The cold receptor. *Amer. J. Psychol.*, 1929, **40**, 353–371.

RANSON, S. W., & CLARKE, S. H. *The anatomy of the nervous system.* (10th ed.) Philadelphia: Saunders, 1959.

ROSE, J. E., & MONTCASTLE, V. B. Touch and kinesthesis. In J. Field (Ed.), *Handbook of physiology.* Vol. 1. Washington, D. C.: American Physiological Society, 1959.

SINCLAIR, D., WEDDELL, G., & ZANDER, E. The relationship of cutaneous sensibility to neurohistology in the human pinna. *J. Anat.*, 1952, **86**, 402–411.

SPERRY, R. W. Neuronal specificity. In P. Weiss (Ed.), *Genetic neurology.* Chicago: Univer. Chicago Press, 1950.

STRUGHOLD, H. Die Topograhie des Kältesinnes in der Mundhöhle. *Z. Biol.*, 1925, **83**, 515–534.

STRUGHOLD, H., & KARBE, M. Die Topographie des Kältesinnes auf Cornea and Conjunctiva. *Z. Biol.*, 1925, **83**, 189–200.

VENDRIK, A. J. H., & VOS, J. J. Comparison of the stimulation of the warmth sense organ by microwave and infrared. *J. appl. Physiol.*, 1958, **13**, 435–444.

VON EULER, C. Selective responses to thermal stimulation of mammalian nerves. *Acta physiol. Scand.*, 1947, **14**, Suppl. 45.

VON FREY, M. Beiträge zur Sinnesphysiologie der Haut. *Ber. König. Säch. ges. Wiss., Math-Phys, Leipzig.*, 1895, **47**, 166.

VON FREY, M., & KIESOW, F. Ueber die Function der Tastkörperchen. *Z. Psychol.*, 1899, **20**, 126–163.

WEBER, E. Der Tastsinn und das Gemeingefuhl. *Wagner's Handworterb. Physiol.*, 1846, **3**, 481.

WEDDELL, G. The pattern of cutaneous innervation in relation to cutaneous sensibility. *J. Anat.*, 1941, **75**, 346–367.

WEDDELL, G., & PALLIE, W. Observations of the neurohistology of cutaneous blood vessels. In G. E. Wolstenholme & J. S. Freeman (Eds.), *Peripheral circulation in man.* (A Ciba Foundation Symposium) London: Churchill, 1954.

WEDDELL, G., PALLIE, W., & PALMER, E. The morphology of peripheral nerve terminations in the skin. *Quart. J. micro. Sci.*, 1954, **95**, 483–501.

WEDDELL, G., PALMER, E., & PALLIE, W. Nerve endings in mammalian skin. *Biol. Rev.*, 1955, **30**, 159–195.

WEISS, P. A. The problem of specificity in growth and development. *Yale J. Biol. Med.*, 1947, **19**, 235–278.

WILLIAMS, R., MCGLONE, B., & BAZETT, H. A comparison of the histology and physiology of the sensory end-organs of the prepuce. *Amer. J. med. Sci.*, 1929, **178**, 887–888.

WINKILMANN, R. K. Some sensory nerve endings in the skin. *Arch. Dermatol. Syphilol.*, 1955, **71**, 373–378.

WOOLARD, H. H. The innervation of blood vessels. *Heart*, 1926, **13**, 319–336.

WOOLARD, H. H. Observations on the termination of cutaneous nerves. *Brain*, 1935, **58**, 352–367.

WOOLARD, H. H. Intra-epidermal nerve endings. *J. Anat., Lond.*, 1936, **71**, 54–60.

ZOTTERMAN, Y. Special senses: Thermal receptors. *Annu. Rev. Physiol.*, 1953, **15**, 357–372.

ZOTTERMAN, Y. Thermal sensation. In J. Field (Ed.), *Handbook of physiology.* Vol. 1. Washington, D. C.: American Physiological Society, 1959. Pp. 431–458.

ZWEIFACH, B. W. Structural aspects and hemodynamics of microcirculation in the skin. In S. Rothman (Ed.), *The human integument.* Washington, D. C.: American Association for the Advancement of Science, 1959.

(Received July 21, 1960)

Psychological Review
1962, Vol. 69, No. 1, 34–53

AN ASSOCIATION MODEL FOR RESPONSE AND TRAINING VARIABLES IN PAIRED-ASSOCIATE LEARNING [1]

GORDON H. BOWER

Stanford University

This paper reports an attempt to apply a simple association learning model to the analysis of the influence of response and training variables on paired-associate learning. The issues under investigation are old ones but have not been resolved satisfactorily in the past. With the aid of the elementary learning model, the problems are posed clearly and, the data willing, adequately resolved by the use of a few simple and intuitively compelling assumptions about learning.

The first problem that led to this investigation concerns the relationship betwen the number of response alternatives (N) and error rate in paired-associate learning. Experimental results (Noble, 1955; Riley, 1952) are in agreement in showing that the number of errors subjects make before reaching some criterion of learning is greater the larger the number of response alternatives. There is little agreement, however, as to the interpretation of this fact.

Two possible factors could be involved. First, the effectiveness of a reinforced trial in increasing performance (i.e., the learning rate constant) may be influenced by N; and second, N may influence the probability of being correct by sheer guessing on items that are yet unlearned. It is a reasonably safe assumption that N has the second effect on chance guessing. Previous data are unclear on whether N also influences the first factor, the effectiveness of a reinforced trial. This is not an easy question to answer since it is difficult to separate the effects of these two factors in the data. Guessing occurs only on unlearned items but there is no way to tell by direct observation just how many items have been learned and how many have been guessed correctly on any given trial. Moreover, there appears to be little hope that more refined or ingenious experimental procedures will enable us to unconfound these two factors so that we may crucially test the hypothesis that N affects the learning rate.

This is the type of situation in which a theoretical model of learning can make a strategic contribution. Indeed, without the aid of some formal model of the learning process, the question of the effect of N can neither be posed clearly nor answered clearly. With the aid of a theory one can make suitable allowance for the guessing factor and thus make an assessment of the contribution of learning and guessing at every stage of the experiment.

The model to be presented has been formally treated in more detail in a previous paper (Bower, 1961). Here the theory will be presented informally and only those implications relevant to the present discussion will be introduced. For a more formal statement of the axioms and theorems, the reader may consult the prior report. The basic notion of the model is the assumption that each stimulus item and its correct response become associated on an all-or-none basis. Considering a single item, it can be in either of

[1] This investigation was supported in part by a grant, M-3849, from the National Institute of Health, United States Public Health Service.

two states at the beginning of each trial: conditioned to its correct response or not conditioned. If the item is conditioned at the beginning of a trial, then the correct response occurs. If the item is not conditioned, then the probability of the correct response depends somewhat upon the experimental procedure. In experiments by the writer, the subjects were told the N responses (Integers 1, 2, . . . N) available to them and were told to respond on every trial regardless of whether they knew the correct answer. If the N numbers occur equally often as the to-be-learned responses to the items, then the probability that the subject will guess correctly on an unlearned item is $1/N$ on the simplest assumptions;[2] correspondingly, his probability of guessing incorrectly is $1 - 1/N$. The following discussion of the model is oriented specifically towards such an experimental procedure.

[2] The arguments to be developed apply only to those conditions in which the response alternatives are immediately available to the subject and he is permitted time to give a relevant response on each trial before the terminal event (e.g., reinforcement) occurs. This is the procedure most familiar to those who run animal subjects in choice situations. A more frequent procedure in human learning experiments has been to give the subject a fixed time interval in which he may respond, and the terminal event (e.g., correct response in paired associates, next item in a serial list) follows after that time interval regardless of whether the subject makes a relevant response. To remark on the sociology of experimenters, there seems to be an empirical correlation between the use of experimenter vs. subject-controlled exposure time and whether the relevant responses per se (as well as the S-R associations) must be learned in the experimental situation: when the responses must be learned, experimenters usually control exposure time; when the response alternatives are immediately available, either via instruction or construction, the subject is allowed to control exposure time. Clearly, however, there is no necessary entailment between these procedural aspects, and one may expect the empty cells of the 2×2 contingency table to be filled by future experimentation.

With the theory formulated in this way, one cannot uniquely specify the subject's state of conditioning on a given item from the knowledge that he made a correct response, since this correct response may have come about by guessing on an unlearned item. Thus, the theory does not imply that the first correct response will be followed by correct responses on all subsequent trials. However, if the subject makes an error, then we can make a determinate inference that the item was not conditioned at the beginning of the trial.

We have introduced the notions of the two states of conditioning and the probability of the correct response given the item's state of conditioning at the beginning of each trial. We assume that each item in initially unconditioned and that the effect of successive reinforced trials is to provide repeated opportunities for the item to become conditioned. The single parameter of the theory is the learning rate constant (c) which represents the probability that an unconditioned item becomes conditioned as the result of a single reinforced trial. The probability that a reinforced trial fails to condition the correct response to an unlearned item is $1 - c$. The probability that an item is still not conditioned after n reinforced trials is $(1 - c)^n$. If c is larger than zero, then this probability approaches zero as n becomes large; that is, given a large number of reinforced trials, it it certain that the item will become conditioned on some one of these reinforced trials. When the item becomes conditioned, the probability of a correct response jumps from $1/N$ up to 1.

From the considerations above we

may obtain an expression for q_n, the probability that the subject responds incorrectly to a given item on Trial n of the experiment. To obtain q_n we note that the probability that the item has failed to be conditioned during the preceding $n-1$ trials is $(1-c)^{n-1}$; and if the item is not conditioned then the probability of guessing incorrectly is $1-1/N$. Hence, the probability of an error on Trial n is given by the product of these two factors:

$$q_n = (1 - 1/N)(1 - c)^{n-1} \quad [1]$$

Using this elementary association model, the first question of this investigation can be restated in a clear manner: Is c independent of N? If c is independent of N, then, given an estimate of c obtained from one group, we should be able to predict in advance the performance of other groups trained with differing numbers of response alternatives by simply adjusting the N factor in Equation 1. To the extent that these free predictions are accurate, we would have good evidence for the assertion that N affects guessing probabilities but has no influence on learning rate per se.

The second question prompting this investigation concerns the relation between several different training conditions and performance in paired-associate learning. The two major conditions of training studied to date are the correction and noncorrection conditions. The more frequently used condition is the correction procedure in which subject is informed of the correct response on every trial regardless of whether he responds correctly. In noncorrection training, the subject is told right or wrong depending on whether his response is correct or incorrect. In the following, we will use the word "reinforcement" in the general sense intended by Estes (1960), viz., as operations exerting certain general quantitative effects upon response probabilities. In paired-associate experiments, the reinforcing operation is that of informing subjects of the correct response or in some way ensuring that this response is the last one to occur in the presence of a stimulus item. With this interpretation, then, on every trial under a correction procedure the correct response is reinforced. The noncorrection training procedure differs in that the correct response is reinforced only on those trials when it occurs; since all incorrect responses are followed by the experimenter simply saying "Wrong," no explicit reinforcement operation is involved. We will refer to such trials as nonreinforced trials.

Within the model, an account of the noncorrection procedure takes the form of specific assumptions about the formation of associations following reinforced and nonreinforced trials. Considerations of parsimony lead us to assume that the effect on conditioning of a reinforced trial is the same whether the reinforcement occurs in the context of a correction or noncorrection procedure, viz., the probability that an unlearned item becomes conditioned to its reinforced response is c, the same constant as in the case of the correction procedure. A new assumption that is required concerns the effect of a nonreinforced trial, when the experimenter says "Wrong" following an error. The assumption that has been made elsewhere (Thorndike, 1932) and that is made here also, is that a nonreinforced trial in this situation results in no net change in the probability of a correct response. This assumption is not offered as a universal interpretation of nonreinforcement; however, within the specific situation to which the model will be applied (i.e., the alternative responses are equiprobable initially and a num-

ber of different S-R events occur between successive presentations of any given stimulus) there is some basis for thinking that the no change assumption may be approximately correct.

From these assumptions about the noncorrection procedure, a difference equation can be derived expressing the change in the average probability of an error from Trial n to Trial $n+1$:

$$q_{n+1} = q_n[q_n] + (1 - q_n)[(1 - c)q_n] \quad [2]$$

The first term on the right-hand side of Equation 2 exhibits the assumption of no change (i.e., $q_{n+1} = q_n$) when the subject makes an error on Trial n, with probability q_n; the second term shows that if a success occurs (with probability $1 - q_n$), then with probability c the item becomes conditioned (if it is not already) and with probability $1 - c$ the reinforcement was ineffective and no change occurs on that particular trial. An explicit solution of Equation 2 is not available. However, Miller and McGill (1952; see also Bush & Mosteller, 1955, p. 181), have derived the following recurrence relation for Equation 2 (the constants have been changed appropriately):

$$q_n = (1 - 1/N)\,[1 - (1 - c)^{n-1}]q_{n-1} + (1 - 1/N)(1 - c)^{n-1} \quad [3]$$

Equation 3 can be used to compute successive values of the average error probability once N and c are known. In general, the predicted learning curve is **S** shaped, a feature which is consistent with the expectation of a low rate of improvement early in training when the subject is trying to discover the correct response. The point of inflection of the **S** curve will be positively related to N. These features seem in qualitative agreement with noncorrection results reported by Noble (1955), although his experimental procedure was more complex than the one under present consideration.

At this point in the analysis, Experiment I was carried out with four groups of subjects learning the same list of 10 paired associates. The variables were the number of response alternatives (the first three or first eight integers) and correction vs. noncorrection training procedure. For brevity, the four groups will be referred to by the symbols 3-C, 8-C, 3-NC and 8-NC, where 3 or 8 represent the number of responses, and C and NC designate correction and noncorrection training, respectively. A more extensive discussion of the procedure and results of Experiment I will be deferred till later; it will be sufficient here to note one critical result of Experiment I since it was this fact that led to Experiment II. The fact was that strong evidence was obtained to support the assumption of no change in success probability following a nonreinforced trial in the noncorrection training groups; that is, with either three or eight response alternatives, saying "Wrong" followed a subject's incorrect response had no effect on his probability of success on the next trial.

This no change result supports our earlier assumption about nonreinforced trials, but how are we to understand it? On rational grounds, one might expect that subjects would tend to eliminate a response which was followed by "Wrong." For example, if subjects in the 3-NC group tended to eliminate their first erroneous response, then the probability of the correct response on the next trial should be around one-half or at least greater than one-third. However, the results of Experiment I clearly showed that this increase did not occur, and

we seek some explanation for why it did not occur.

An explanation may be found perhaps by attending to the responses evoked in the subject by saying to him "Wrong" after he has responded, e.g., "three." The reason for this inquiry is that, holding to a strict contiguity interpretation for the formation of association, it is these terminal reactions to the stimulus which have an opportunity to become conditioned. Often these implicit reactions to "Wrong" are primarily emotional; in this regard, it may be reported that subjects frequently volunteered the information, "I know I've been getting that item wrong, but I can't remember what number I said last time." Clearly, if the subjects' implicit reactions are primarily emotional, then we may expect no change in the recorded response probabilities on the next trial. However, if subjects implicitly react to "three is wrong" with the response "one or two is correct," when the contiguity interpretation would imply that the probability of response "three" would decrease and the probability of responses "one" and/or "two" would correspondingly increase.

According to this interpretation, the noncorrection procedures in Experiment I did not insure responses of the form, e.g., "one or two is correct" after the experimenter said "Wrong" to "three." For one of the groups in Experiment II, conditions were arranged to insure the occurrence of responses in this form following errors. There were three response alternatives and subjects were instructed that one and only one number was correct for each nonsense syllable. If, for example, the subject responded with "three," instead of saying "Wrong," the experimenter said "one or two is correct." It should be noted that, in a formal sense, the subject gains no more information about the correct response with this procedure than when the experimenter says "Wrong," as in the noncorrection procedure. Comparisons of the results of this procedure with the standard noncorrection procedure are given below.

The second condition run in Experiment II was aimed at a slightly different question. The question was whether there is some more basic way of specifying reinforcement contingencies in this situation rather than merely listing them, correction and noncorrection. The formulation proposed here is that this basic variable is the degree to which the experimenter specifies the correct response following an error by the subject; to describe it in another way, the variable is the size of the subset of alternatives within which the one correct alternative is said to lie. The correction and noncorrection procedures occupy the extreme poles of this dimension; in the correction procedure, the experimenter uniquely specifies the correct response; in the noncorrection procedure, when the experimenter says "Wrong," he implicitly specifies "one of the other $N-1$ alternatives is correct." [3] To illustrate the construction of intermediate values of this training dimension, suppose there were eight response alternatives and that the subject responds incorrectly with "one" to a given item; then the experimenter might say "three or six is correct," or "six or eight or two or five is correct," and so on. For the second group in Experiment II, with eight responses, the experimenter said two numbers following errors. The subjects were instructed that one of the numbers was correct and that the other number was a distractor, but there were no cues

[3] However, see the related discussion above concerning the subject's reactions to "Wrong."

as to which number was the correct one. Of course, if the subject gave the correct number, the experimenter indicated this to him.

It is clear that such a continuum of training conditions can be constructed. It is also clear that variations in this training variable should produce graded variations in performance intermediate between the two extremes produced by the correction and noncorrection procedures. To describe some of the factors involved in this prediction, consider the reasons for expecting the partial correction group listed above (call it the 8-P group) to make fewer errors than the 8-NC group from Experiment I.

First, a subject in the 8-P condition is expected to make fewer errors before his first correct response because there is at least some likelihood that the correct association will be formed following an error. Specifically, we assume that before the first correct response the probability that the item becomes conditioned following an error is $c/2$, where c is the same learning constant as before. In general, if the subject is told k alternatives, one of which is correct, then we assume that the probability of the correct association being formed is c/k on each trial before the first success. This formulation is equivalent to assuming that the subject selects at random one of the k possibly correct responses to rehearse on a given trial, with probability $1/k$ he selects the correct response, and with probability c rehearsal of this correct response results in conditioning.

The preceding analysis applies on trials before the first correct response occurs. After the subject has been once informed of the correct response, his probability of recognizing it among the two numbers the experimenter says (following a subsequent error, if any) will be greater than one-half. If the subject recognizes the correct number and then rehearses it, the effect is much the same as a reinforced trial. In general, we may let r represent this recognition probability following the first correct response; because of implicit rehearsal, r is also the probability of a reinforced trial on incorrect trials following the first correct response. Specifically, we assume for the present case that r is unity; that is, following the first correct response, we assume that subject can recognize with Probability 1 the correct number among the two numbers that the experimenter says following an error. If the subject recognizes the correct response and rehearses it, then with probability c the conditioned association is formed.

To summarize the discussion for the partial correction groups, we have assumed that prior to the occurrence of the first correct response the probability of conditioning is $c/2$ on each trial; on the trial of the first correct response and those trials following the probability of conditioning is c. In a later section we suggest a number of ways to test the details of these assumptions.

In the experiments to be described there were six independent groups; if the model and theoretical assumptions are correct, then the data from the six groups can be reproduced after estimating the constant c from one of the groups selected arbitrarily. One can easily recognize the advantages of casting our theoretical assumptions in explicit form within such a model. Not only can we clearly pose the questions of how N and its interactions with training conditions affect error rate, but we have also developed a conceptual framework within which it is possible to get an answer to these questions. The basic learning parameter,

c, plays a central role in the theory and is not just a curve-fitting constant; for a homogeneous population of subjects with the same learning materials, the underlying theory constrains c to be the same for all conditions. If indeed such parameter invariance obtains (i.e., if c can be transposed from one condition to the others), then a strong appeal can be made for the simplicity and power of the underlying theory within which c derives its meaning.

THE EXPERIMENTS[4]

The subjects, 88 Stanford undergraduates, were required to learn a list of 10 paired associates to a criterion of two successive correct runs through the entire list. The stimulus items were nonsense syllables of 0–5% association value chosen for low intralist similarity. The responses were the first three or first eight integers, paired randomly with the stimuli for each subject. With three response alternatives, two of the three responses were used for three stimuli each and the third response, selected randomly for each subject, was paired with the remaining four stimulus items. With eight response alternatives, all eight numbers were first paired with eight randomly selected syllables for each subject, and then the remaining two syllables were assigned to different but randomly selected numbers for each subject.

The subjects, run individually, were instructed to learn a list of 10 nonsense syllables and their associated number responses, the response alternatives being either one, two, or three (or one, two, . . . , seven, or eight), and that the syllables and numbers had been paired off randomly. They were asked to respond within 2 or 3 seconds after the stimulus card was shown, and to guess a number if they didn't know the answer. The result of this practice is that the exposure time cannot be specified exactly; however, it was noted that the subjects followed instructions and responded within 2 or 3 seconds in the large majority of cases.

For all three training conditions, if the subject gave the correct number to a card, the experimenter repeated it, e.g., "three is correct." The three conditions differed on trials when the subject responded incorrectly: in the correction conditions, the experimenter said the correct number; in the noncorrection conditions, the experimenter said "Wrong" following errors; and in the partial correction condition, the experimenter said two numbers following errors, e.g., "one or three is correct." The subjects in the partial conditions were informed that only one of the numbers the experimenter said would be correct and the other number was a distractor (chosen at random) and that half the time the correct number would be said first and half the time it would be said second. The 10 stimulus cards were shuffled for 15 seconds between runs through the deck. The entire list was repeated until the subject went through two consecutive cycles without any errors.

The numbers of subjects in each group were 14 in the 3-C, 8-C, and 8-NC groups; 15 in the 3-P and 8-P groups; and 16 in the 3-NC group. Assuming that each item and each subject may be characterized by the same value of c, the data consist of 140, 150, or 160 sequences of correct and incorrect responses for the various experimental conditions. Before the model could be evaluated, the learning rate constant, c, had to be estimated. The data from the 3-C group were selected for this purpose;

[4] This experiment was carried out with the assistance of Takao Umemoto.

the least squares estimate of $c = .218$ was obtained by fitting the learning curve (i.e., percentage correct vs. trials) for this group.

The first predictions concern the average total errors per item before achieving the learning criterion. The predictions for the correction and noncorrection groups were obtained by summing Equations 1 and 3, respectively, over trials. The predictions for the partial groups were obtained by calculating the average errors before the first success (using $c/2$ as the rate constant) and the average errors following the first success (using c as the rate constant) and then adding these two numbers. The observed means (M) and standard errors of the means (σ_M) are shown in Table 1 along with the predicted means (P).

It is clear from Table 1 that the theory performs adequately in predicting average total errors using the single estimate of c. In all six cases the predicted value falls within one standard error of the observed mean.[5] The differential effect of N upon error rate is amplified as one proceeds from the correction through the partial to the noncorrection training procedures.

A second indication of the goodness of fit of the theory is shown graphically in Figures 1 and 2 which present observed and predicted values for cumulative errors as a function of trials. The smooth curves connect the predicted values while the unconnected points are the empirical values. The predicted values for the correction and noncorrection conditions were obtained

TABLE 1

AVERAGE TOTAL NUMBER OF ERRORS PER ITEM OBSERVED AND PREDICTED

Training condition	Response alternatives 3	8
Correction	$M = 2.94$ $P = 3.03$ $\sigma_M = .19$	$M = 3.95$ $P = 4.00$ $\sigma_M = .25$
Partial	$M = 3.74$ $P = 3.88$ $\sigma_M = .21$	$M = 5.95$ $P = 5.80$ $\sigma_M = .35$
Noncorrection	$M = 5.65$ $P = 5.59$ $\sigma_M = .34$	$M = 9.42$ $P = 9.54$ $\sigma_M = .61$

by cumulating successive values of error probabilities calculated from Equations 1 and 3, respectively. The theoretical curve for the partial correction conditions is not expressible in a simple equation but successive values can be obtained. The first several values will be derived here to illustrate the procedure. Define Q_n as the probability that the item is not yet conditioned by the beginning of Trial n. The probability of an error on Trial n would then be $(1 - 1/N)Q_n$. The first several values of Q_n for the partial correction conditions are:

by assumption

$$Q_1 = 1$$

$$Q_2 = \frac{1}{N}(1-c) + \left(1 - \frac{1}{N}\right) \times \left(1 - \frac{c}{2}\right)$$

$$Q_3 = \frac{1}{N}(1-c)^2 + \left(1 - \frac{1}{N}\right) \times \left(1 - \frac{c}{2}\right)\frac{1}{N}(1-c) + \left(1 - \frac{1}{N}\right)^2\left(1 - \frac{c}{2}\right)^2 \qquad [4]$$

[5] The standard errors for the 3-C and 8-C groups can be predicted by the theory and turn out to be somewhat larger than the observed values (predicted values were .24 and .30 for 3-C and 8-C groups, respectively). Deriving comparable variance predictions for the partial and noncorrection conditions presents complex mathematical problems which have not been solved to date.

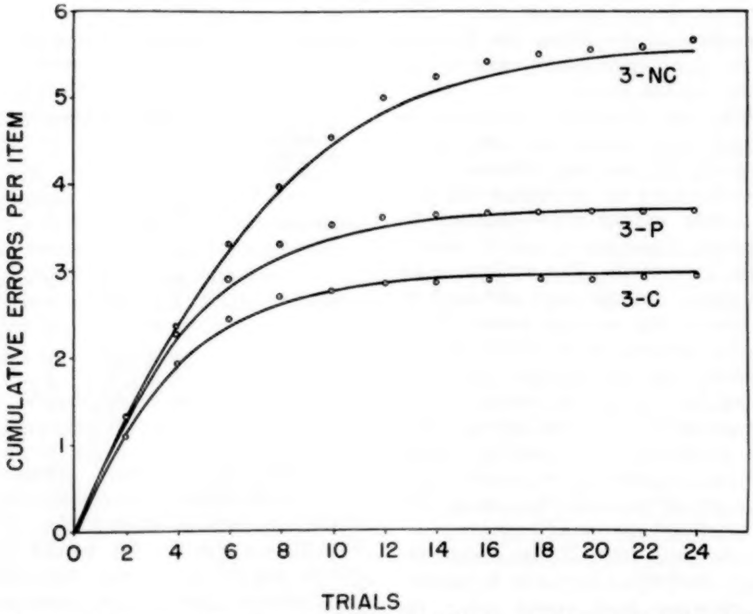

FIG. 1. Average cumulative errors per item plotted as a function of trials for subjects having three response alternatives. (The smooth curves represent predicted values; the dots near a curve represent the corresponding observed values.)

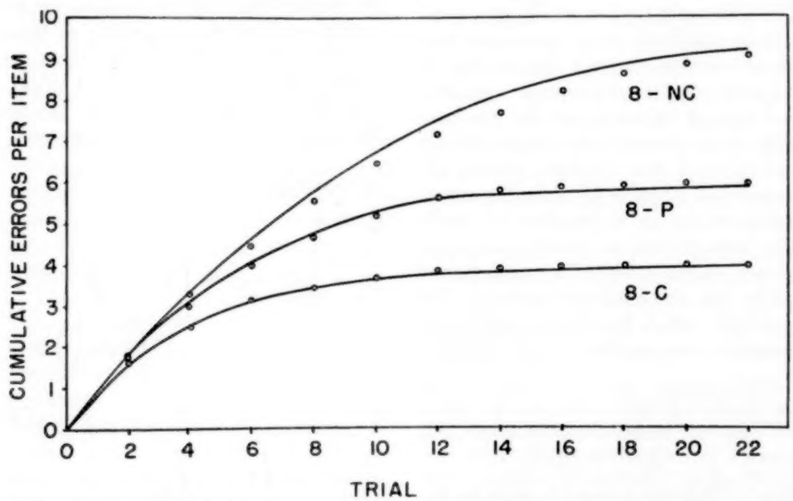

FIG. 2. Average cumulative errors per item plotted against trials for subjects having eight response alternatives. (The smooth curves represent predicted values; the dots near a curve are the corresponding observed values.)

The value of Q_2 is the sum of two probabilities: (*a*) that a correct guess occurs on Trial 1 but conditioning failed to occur with probability $1 - c$, and (*b*) that subject guessed incorrectly on Trial 1 and fails to condition the correct response with probability $1 - c/2$. The value of Q_3 is the sum of the joint probabilities of three events: (*a*) correct on Trial 1, fails to condition on Trials 1 and 2; (*b*) error on Trial 1, success on Trial 2, fails to condition after either trial; (*c*) errors on both trials, not conditioned on either trial. The general rule for obtaining Q_n is that prior to the first correct response the conditioning probability is $c/2$ and that on the trial of the first correct response and thereafter the conditioning probability is c. By multiplying these Q_n values by $1 - 1/N$ and summing them, the predicted curves for the partial correction conditions are obtained. As Figures 1 and 2 show, for five of the six groups the fit of predicted to observed values is satisfactory; the correspondence to the 8-NC points would not be impressive were it not for the fact that the same c constant was used in generating all the curves under appropriate boundary conditions.

Equations 2 and 3 for the noncorrection conditions were derived with the assumption that no change resulted from nonreinforced trials. To the extent that these equations adequately describe performance in the noncorrection conditions, the assumption of no change is supported. However, additional data bearing upon this assumption may be obtained by analysis of the early trials before learning has occurred. Beginning with the first trial, consider the sequence of probabilities of a correct response following 0, 1, 2, . . . consecutive errors. As soon as a subject responds correctly to a given item, that item is

TABLE 2

PROBABILITY OF A SUCCESS FOLLOWING n CONSECUTIVE ERRORS FROM THE FIRST TRIAL FOR THE NONCORRECTION GROUPS

n	3-NC	8-NC
0	.34	.12
1	.31	.11
2	.28	.14
3	.31	.17
4	.31	.13
5	.37	.16

dropped from later computations. If the randomization of responses to stimuli has been effective, then the first value of this series should be close to $1/N$, the probability of guessing the correct number by chance. Moreover, if the assumption of no change is correct, then the probabilities of a success following 1, 2, 3, . . . consecutive errors should remain around $1/N$ for the noncorrection groups, deviating from this value only by random sampling fluctuations. The results of this analysis are shown in Table 2 for the noncorrection groups. It should be realized that going down the column these estimates become progressively less reliable because the number of cases involved (beginning with 160 or 140) decreases. These series of estimates for both three and eight response alternatives reveal no significant trends away from the first value. Hence, the results provide additional support for the assumption of no change following nonreinforced trials in the noncorrection conditions.

An analysis related to that above can be given for the correction and partial correction groups. Because the success rate was higher for these groups, there was insufficient data to obtain reliable estimates of single-trial probabilities such as shown in Table 2 for the noncorrection groups. How-

TABLE 3

PREDICTED AND OBSERVED CONDITIONAL PROBABILITIES OF SUCCESS FOLLOWING AN ERROR ON THE PRECEDING TRIAL

Training condition	Response alternatives			
	3		8	
	Observe	Predict	Observe	Predict
Correction	.48	.48	.35	.35
Partial				
Before first success	.35	.35	.21	.22
After first success	.46	.48	.38	.33

ever, this difficulty may be circumvented by considering over all trials the probability of a success given an error on the preceding trial. For the correction group, this probability should be a constant, $c + (1 - c)p_o$ where p_o is the probability of guessing correctly. For the partial correction conditions, the probability of a success following an error should be $c/2 + (1 - c/2)p_o$ before the first success and should be $c + (1 - c)p_o$ after the first success. Such predictions are very sensitive to small changes in p_o. Our efforts to randomly assign responses to stimuli were directed towards making p_o be close to $1/N$ in value. However, with even a moderate-sized sample, small deviations of the actual probability of correct guessing inevitably occur. Hence, a fairer test of the model (rather than a test of the randomization procedure) is provided by predictions based on empirical estimates of p_o. Using the first-trial estimates of p_o (which on the whole were close to $1/N$), the predicted probabilities of success following an error were obtained and are compared with the observed values in Table 3. The correspondence between predicted and observed values is quite satisfactory; we may note that the rather complex assumptions about conditioning in the partial correction conditions receive substantial support from these data.

Another way to test the assumptions of the model involves predicting the number of errors before the first success. Since for the partial and noncorrection groups the stochastic process is not complicated prior to their first success, it is possible to predict both the mean and variance of this statistic for these groups. Define F_1 to be the number of errors before the first success, and p_o to be the probability of a correct guess. Then it can be shown (Bower, 1961) that the mean and variance of F_1 for the noncorrection groups are:

$$M = \frac{1 - p_o}{p_o}, \quad V = \frac{1 - p_o}{p_o^2} \qquad [5]$$

Similarly, for the correction groups, the mean and variance of F_1 are:

$$M = \frac{1 - p_o}{p_o + c(1 - p_o)}, \qquad [6]$$

$$V = M + M^2(1 - 2c)$$

The expressions for the partial correction groups are the same as in Equation 6 except with c replaced by $c/2$.

The predicted values of the mean and standard error of F_1 for the six groups are shown in Table 4 along with the observed means and standard errors. For the noncorrection groups, p_o was estimated as the average of the values in Table 2 since the theory says those values represent random fluctuations around p_o.

Inspection of Table 4 shows that in

the six cases the predicted mean F_1 is within a standard error of the observed mean, and the predictions of the standard errors are also close to the observed standard errors. In general, the average F_1 increases with N and the differential effect of N is amplified under the partial and noncorrection training conditions.

It should be noted that at an empirical level N has its primary influence upon errors before the first success, and has practically no influence upon errors following the first success. The reader may convince himself of this fact by subtracting the mean F_1 values in Table 4 from the mean total error values in Table 1. At first glance, this observation seems to invalidate the model by showing that N only influences the discovery stage but not the fixation stage of learning after the first correct response has occurred. In fact, however, theoretical predictions of average errors following the first success correspond closely to the observed values; of course, this must be so since the theory predicts F_1 and total errors reasonably well. The reason we expect errors following the first success to be relatively constant over N is that, according to the theory, a sampling bias is involved in comparing the three and eight alternative conditions at this point. Errors following the first success will occur only for items whose first correct response came about by guessing, and for these guessed correct items we expect more errors with eight than with three response alternatives. Let g_N represent the probability that with N alternatives the first correct response comes about by guessing and let e_N represent the average number of subsequent errors for an item whose first correct response occurred by guessing. Then the expression for the average number of errors following the first correct response (call it w_N) is:

$$w_N = g_N e_N + (1 - g_N)\cdot 0 = g_N e_N \quad [7]$$

That is, a proportion g_N of first correct responses come about by guessing and we expect an average of e_N more errors for these items; a proportion $1 - g_N$ of the first correct responses come about by conditioning on the preceding trial and for these items we expect zero subsequent errors. According to the theory, e_8 will be larger than e_3; however, g_8 will be smaller than g_3; that is, given a first correct response, it is more likely to have occurred by guessing when N is three than when N is eight. Therefore, the fact that the products g_3e_3 and g_8e_8 are approximately equal is neither surprising nor a refutation of the theory.

DISCUSSION

The overall results provide strong support for the validity of the model and the specific hypotheses about ex-

TABLE 4

OBSERVED AND PREDICTED MEAN AND STANDARD ERROR OF THE NUMBER OF ERRORS BEFORE THE FIRST SUCCESS

Training condition	Response alternatives: 3	Response alternatives: 8
Correction	M = 1.33 F = 1.39 σ_M = .092 σ_F = .125	M = 2.25 F = 2.37 σ_M = .185 σ_F = .180
Partial	M = 2.12 F = 2.16 σ_M = .192 σ_F = .191	M = 4.15 F = 4.02 σ_M = .289 σ_F = .329
Noncorrection	M = 2.17 F = 2.15 σ_M = .201 σ_F = .206	M = 6.20 F = 6.38 σ_M = .511 σ_F = .558

perimental variables. The data of six independent groups were adequately reproduced using a single estimate of the learning rate constant obtained from one of these groups. The fact that in 24 independent predictions the model came close to the data is sufficient justification for exploring further consequences of the theory.

The initial problem that led to this investigation was whether the number of response alternatives could be shown to affect learning rate on reinforced trials in addition to contributing to differential error probabilities due to guessing on unlearned items. The data supported the assumption that learning rate on reinforced trials was a constant independent of the number of response alternatives, and that the effect of N upon error rate could be attributed to differential guessing probabilities on unlearned items. It should be noted that these conclusions are valid only for the experimental procedure used here in which the response alternatives are immediately available to the subject. For alternative procedures in which the subject is required to learn the responses (e.g., three or eight nonsense syllables) as well as the S-R associations, a number of complicating factors enter to obscure the picture and prevent resolution of the basic problem.

The second problem of this study was to account for the effect on performance of several training procedures. One may conceive of the correction, partial correction, and noncorrection procedures as varying in the degree to which the correct response is specified following an incorrect response by the subject. This variable in turn determines the probability that the correct response is reinforced following incorrect responses. Thus, in the correction condition, the reinforcement probability is 1 on every trial; under partial correction, the reinforcement probability is $1/k$ before the first correct response and is essentially 1 afterwards due to a high recognition probability; under noncorrection, the probability that the correct response is reinforced is essentially zero following errors. The data supported these hypotheses and gave additional support to the assumption that when the correct response was reinforced, the probability of conditioning was a constant, c, the same for all three conditions.

In addition to confirming these specific assumptions about N and training conditions, the results lend support to the learning model in which these assumptions are embedded. In this one-element model, the probability of the correct response can have only two values, $1/N$ or 1. The stimulus element begins in the unconditioned state and each reinforced trial provides an opportunity for the element to become associated in all-or-none fashion with the correct response. When conditioning finally occurs, response probability jumps from $1/N$ to 1. Discontinuous learning theories have been contrasted frequently with "response strength" theories which assume that an associative factor (habit strength in Hull's theory, proportion of conditioned elements in Estes' linear model) increases in a cumulative manner with successive reinforcements. According to a strength theory, response probability make take on a large (possibly infinite) number of values ranging from $1/N$ up to 1. It is frequently difficult to distinguish this theory from the one-element model since they predict the same average learning curve over a group of subjects and items. Indeed, the present results on average errors could be fit about equally well

by the linear model. Since the discussion has been oriented towards the one-element model, perhaps it would be appropriate to record a few extra results which favor this model. For these purposes, let us consider two sequential statistics for which the two models deliver qualitatively different predictions.

The first statistic is the average number of errors (to perfect learning) following an error that occurs on Trial n of the experiment. The one-element model makes the rather counterintuitive prediction that this average number of subsequent errors is a constant independent of the trial number on which the leading error occurs. Thus, if we observe an error on Trial 10 we predict the same number of subsequent errors as if we had observed that error on Trial 1. The point of the matter is that when an error occurs on Trial n we know that the item was not conditioned before Trial n, and we can set the clock back to Trial 1 as far as the model is concerned in predicting future errors on that item. In contrast, the associative strength approach (specifically, the linear model) predicts that the number of errors following an error on Trial n is a decreasing function of n; that is, the greater the number of reinforced trials before a particular error, the higher the associative strength at that time, and hence the fewer the number of subsequent errors before perfect learning.

To obtain a sizable sample on which to test this critical point, data from 48 subjects in another 10-item paired-associate experiment[6] (with $N = 2$) were pooled with the 3-C and 8-C groups of the present experiment. The varying N's will not affect the

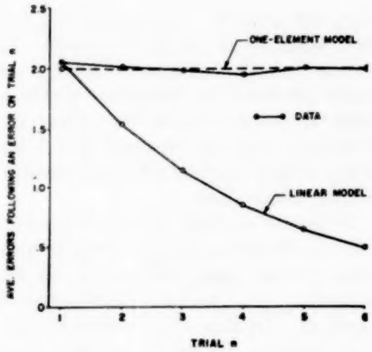

FIG. 3. Paired-associate data: average errors following an error that occurs on Trial n, for $n = 1, 2, \ldots, 6$. (Data points are filled dots; predictions of linear and one-element models are indicated.)

results on constancy vs. monotone decreasing aspects of errors following an error on Trial n. Using this pool of 760 response sequences (10 items for each of 76 subjects), the distributions of the number of errors following an error occurring on Trial 1, on Trial 2, . . . , on Trial 6 were obtained. The analysis was not carried beyond Trial 6 since the number of cases involved was decreasing so that estimates of means would be more unreliable. The estimates of the average number of errors following an error on Trial n are shown in Figure 3.

These data clearly favor the constancy prediction of the one-element model. Although the estimates fluctuate somewhat this can be attributed to sampling variability; even the largest difference (2.05 vs. 1.95) does not approach statistical significance ($t = 0.49$). Also in Figure 3 is shown the rough order of magnitude of the numbers to be expected from the linear model. These predicted numbers are not exact because of different learning rates in the two experiments that were pooled. How-

[6] These experiments were carried out with the assistance of Norman Karns and James Colloran.

ever, the average c value was .25 and this value was used to obtain the values on the graph. Clearly the curve predicted by the linear model is quite discrepant from the data; accordingly, the one-element model appears to give a more adequate description of these data.

A second example of this constancy effect is shown in Figure 4 which presents similar data collected in an experiment on verbal discrimination learning. In this experiment 34 subjects learned 20 items, an item consisting of two different nonsense syllables printed on a card. The subject's task was to read off each card that syllable arbitrarily designated as correct for that item; the subject repeatedly went through the items in random order until he achieved two consecutive errorless trials with the entire list. Since response learning per se is not involved in this situation, the one-element model was expected to apply. Since there were two response alternatives on each card, we make the natural assumption that the chance probability that subject selects the correct syllable is one-half. Detailed comparisons of predicted and observed statistics indicated excellent fit of the model; the results in Figure 4 are representative of the overall accuracy of the predictions for this set of data. An additional set of paired-associate learning data providing a comparison of the linear and one-element models has been reported in a previous paper (Bower, 1961); there again the one-element model was clearly superior to the linear model in predicting details of the data.

A second measure suggested by Estes (1960) for differentiating the one-element and linear models con-

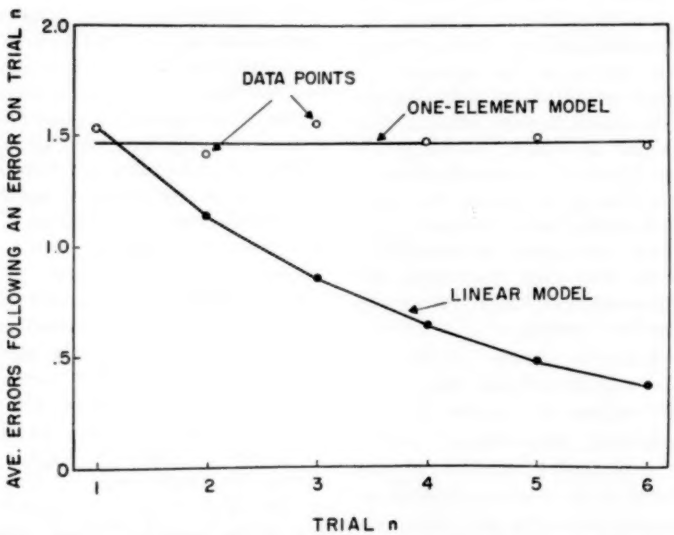

FIG. 4. Verbal discrimination data: average errors following an error that occurs on Trial n, for $n = 1, 2, \ldots, 6$. (Data points are the dots: predictions of linear and one-element models are indicated.)

cerns the probabilities of repetition and alternation of responses over a series of test trials following training. Consider a miniature experiment consisting of one or two reinforced trials followed by two test trials (T_1 and T_2) on which the subject is not informed whether his response is right or wrong. The comparison differentiating the two theories is the probability of correct on T_2 given a correct and incorrect response, respectively, on T_1. According to the one-element model, the probability of correct following incorrect on T_1 should be around the chance guessing level ($1/N$); the probability of correct following correct on T_1 will be much higher, being somewhat less than one because we may expect some of the correct responses on T_1 to have occurred by guessing. The predictions of these conditional probabilities by the linear model depends upon whether there are substantial subject and/or item differences in learning rate. If for the moment we neglect such differences, as we have in testing the one-element model, then the linear model implies that the conditional probabilities of success following either a success or failure on T_1 will be equal.

Two experiments will be reported to this point. The first, performed with the assistance of Sharon Gadberry, consisted of 36 subjects learning two 10-item lists (nonsense syllables), the responses being the integers 1 to 10. Following two presentations of each S-R pair, the subjects received two test trials with the stimulus member of each pair. The pooled results of this experiment are shown in the first row of Table 5. Starting with 720 cases (36 subjects on two 10-item lists), the probability of correct on T_1 was .586; of those correct, 89.6% were

TABLE 5

RESULTS OF TWO EXPERIMENTS ON THE PROBABILITY OF CORRECT ON T_2 FOLLOWING CORRECT AND FAILURE ON T_1

Experiment	Number of cases	Percentage correct on T_1	Percentage correct after correct	Percentage correct after failure
I. Nonsense syllables, 10 number responses	720	.586	.896	.163
II. Noun-noun pair.	320	.900	.995	.030

correct on T_2; of those incorrect on T_1, only 16.3% were correct on T_2. These results support the one-element model. Two comments are required here: First, the value of .896 for correct following correct on T_1 is about what would be expected if 10% of the correct responses on T_1 had occurred by guessing; secondly, the value of .163 for correct following incorrect on T_2 is a little higher than the chance level of .100; however, this is likely an artifact of assigning stimuli and responses in one-to-one correspondence so that when k items are learned the probability of guessing correctly on the unlearned items may be close to one in $N - k$ instead of the a priori value of one in N.

The difference between probabilities of success on T_2 following a success and failure on T_1 should be increased above those in the previous study if one requires the responses per se to be learned (e.g., as in word-word pairs). Under these circumstances, the chance level of guessing will be practically zero. Thus, if the subject fails to respond correctly on T_1, the probability of being correct on T_2 is essentially zero; by the same reasoning, if a correct response occurs on T_1 it is most likely to have come about by condi-

tioning so that the probability of correct on T_2 for this item should be essentially unity if forgetting is negligible in this situation. Row two of Table 5 reports the results of an experiment performed under these conditions with the assistance of Judith Slagter. Thirty-two subjects received three presentations on each of 10 noun-noun pairs (e.g., moon-pin) followed by two test trials. The percentage correct on T_1 was substantially higher than in the first experiment with nonsense syllable-number pairs. However, the important points of interest are that the probability of correct on T_2 given a correct on T_1 is essentially unity, whereas the probability of correct following failure on T_1 is nearly zero.

One objection frequently raised to these foregoing comparisons is that if there are substantial subject and/or item differences in learning rates, then arguments based upon the pooled aggregate will tend to favor the one-element result. The point of the objection is this: if the linear model holds, and if differences in learning rates lead to a distribution over subjects of response probabilities, then considering the entire aggregate the conditional probability of a success following a success is expected to be higher than the conditional probability of success following a failure. This happens, according to the argument, because when we conditionalize upon the first success we are selecting predominately those protocols from the upper end of the distribution of response probabilities. However, granting that this argument is sound, it nevertheless helps not at all in rescuing the linear model from such data; it simply shifts the focus of the argument from conditional probabilities to joint probabilities of pairs of responses on T_1 and T_2. For convenience in the following, let p_{10} represent the joint probability of a success on T_1 and a failure on T_2, and let p_{01} be the joint probability of failure on T_1 and success on T_2. It is a simple matter to show that if the linear model holds, then p_{10} should equal p_{01} *regardless* of possible differences between subjects and/or items in learning rates or initial probabilities. Data from four studies previously published by Estes (1960, 1961) show that the required identity of p_{10} and p_{01} fails to appear in any of the four studies. For example, in one study (Estes, 1960, Figure 2, p. 215) the value of p_{10} was .132 while p_{01} was .046; in another study (Figure 6, p. 218) p_{10} was .098 while p_{01} was .003. A variety of ad hoc hypotheses about individual differences to supplement the linear model have been considered in a paper by Estes (1961); the outcome of those investigations was that no simple ad hoc supplements could bring the linear model into correspondence with these elementary data which were collected to test it directly.

The question about individual differences raised by the objection above can be answered by comparison with the variability expected from the model. Suppose this question is considered in the context of a more extended experiment such as that described for the 3-C group in the preceding pages. A summary measure of an individual's performance might be the total errors (T) he makes over the 10 items. If we assume within the model that all subjects and items are characterized by the same learning rate constant, c, then a subject's T score represents the sum of 10 values sampled randomly and independently from a distribution which allegedly is the same for all subjects and items. The theoretical variance of these T

scores may be calculated using our single estimate of c; this predicted variance may then be compared with the variance of the observed T scores by taking their ratio which will be distributed approximately as the F statistic. If the observed variance of T scores is much larger than that predicted, then we would tend to attribute the inflated variance to variations in c over subjects. This test has been applied to several sets of paired-associate learning data collected by the writer, including the data from the 3-C and 8-C groups of the present study for which the variance of T could be predicted. With the population of college students who have served as subjects in these various experiments, this test statistic has always yielded nonsignificant F values; that is, the variance of T between subjects does not differ significantly from what would be expected on the basis of random sampling from the stochastic process assumed by the model to be common to all subjects. A similar statement can be made about analogous tests for differences in item difficulty; this outcome was expected, of course, since precautions were taken in selecting stimuli which appeared to be equal in intralist similarity and association value. Alternative procedures for handling this question regarding subject and/or item differences in the RTT experiment were considered in a paper by Estes, Hopkins, and Crothers (1960).

In concluding this report, it might be appropriate to add a few general comments concerning the issues under discussion lest some misunderstandings arise. First, it is misleading to cast the issues under discussion at the level of all-or-none theories versus response-strength theories of learning; rather the point at issue is whether the appropriate model for the present experimental situation is a two-state or a multistate process, where "state" here refers to a particular value of response probability. The continuous linear model is the limiting case of the class of multistate models. The all-or-none assumption about conditioning with respect to the available stimuli is not a differentiating feature of these two classes of models. Since the early writings of Guthrie (1930) it has been clear that the all-or-none conditioning assumption would imply gradual and cumulative changes in response probabilities provided there is sufficient variability in the stimulus samples from trial to trial. Within the framework of statistical learning theory, the feature differentiating the two classes of models is how many independent stimulus components one must assume to represent adequately the course of learning in this or that experimental situation. The two-state model proposed here for elementary association learning assumes that each item may be represented by a single stimulus element within the model and that this element can be in either of two states. That the number of elements is the critical assumption is illustrated by the fact that the statistics in Figures 3 and 4 and in Table 5 would not discriminate qualitatively between the linear model and a *two*-element model (i.e., where each item is represented by two stimulus components, with a random one being sampled on each trial). For more extensive discussions of small element models the reader may refer to a paper by Estes (1959) or a book by Suppes and Atkinson (1960).

Secondly, it should be recognized that in this report we have demonstrated only that the one-element model adequately describes results

from elementary paired-associate learning experiments. The job of extending the range of applicability of a theory is an empirical project which must proceed piecemeal. In large part, the success in applying the model to a given learning situation will depend upon the simplicity of the situation and the degree of experimental control over stimulus variables. To cite a pertinent illustration: in training a rat to shuttle in response to a buzzer to avoid shock in a Mowrer-Miller shuttle box, conditioning is typically a gradual, "multi-stage" process (cf. Mowrer, 1960, p. 36); however, if experimental conditions are drastically simplified and precautions are taken to eliminate possible sources of interfering responses, then avoidance conditioning is a one-trial, "two-state" process (Maatsch, 1959). An analogous illustration for the eyelid conditioning situation has been reported by Voeks (1954).

In contrast to those situations in which a two-state behavioral process may possibly be achieved through experimental control, there is a large number of common learning situation for which such precise stimulus control probably can not be achieved, and it is unlikely that a simple two-state model could apply. These situations may be characterized generally as ones in which exposure to discriminative stimuli is subject-controlled; these situations range from those in which the critical cues are proprioceptive stimuli from the subject's current behavior, as in motor skill learning, to situations in which selected components of a complex stimulus array control behavior through the mediation of an overt or implicit observing response, as in complex concept learning. One example of this latter class would be that of a rat learning a black-white discrimination in a **T** maze; the stimuli effective at the moment of choice are multiple and vary from trial to trial depending upon the subject's vicarious trial and error behavior. For such situations more elaborate conceptual apparatus (cf. Audley, 1960; Bower, 1959; Spence, 1960) is required to analyze the interaction between observing behavior and instrumental, goal-directed behavior.

REFERENCES

AUDLEY, R. J. A stochastic model for individual choice behavior. *Psychol. Rev.*, 1960, **67**, 1–15.

BOWER, G. H. Choice point behavior. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford Univer. Press, 1959.

BOWER, G. H. Application of a model to paired-associate learning. *Psychometrika,* 1961, **26**, 255–280.

BUSH, R. R., & MOSTELLER, F. *Stochastic models for learning.* New York: Wiley, 1955.

ESTES, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford Univer. Press, 1959.

ESTES, W. K. Learning theory and the new mental chemistry. *Psychol. Rev.*, 1960, **67**, 207–223.

ESTES, W. K. New developments in statistical behavior theory: Differential tests of axioms for associative learning. *Psychometrika,* 1961, **26**, 73–84.

ESTES, W. K., HOPKINS, B. L., & CROTHERS, E. J. All-or-none and conservation effects in the learning and retention of paired associates. *J. exp. Psychol.*, 1960, **60**, 329–339.

GUTHRIE, E. R. Conditioning as a principle of learning. *Psychol. Rev.*, 1930, **37**, 412–428.

MAATSCH, J. L. Learning and fixation after a single shock trial. *J. comp. physiol. Psychol.*, 1959, **52**, 408–410.

MILLER, G. A., & MCGILL, W. J. A statistical description of verbal learning. *Psychometrika,* 1952, **17**, 369–396.

MOWRER, O. H. *Learning theory and behavior.* New York: Wiley, 1960.

NOBLE, C. E. Compound trial-and-error learning as a function of response availability. *J. exp. Psychol.*, 1955, **49**, 93–96.

RILEY, D. A. Rote learning as a function of distribution of practice and the complexity of the situation. *J. exp. Psychol.*, 1952, **43**, 88–95.

SPENCE, K. W. Conceptual models of spatial and non-spatial selective learning. In K. W. Spence (Ed.), *Behavior theory and learning.* Englewood Cliffs, N. J.: Prentice-Hall, 1960.

SUPPES, P., & ATKINSON, R. C. *Markov learning models for multiperson interactions.* Stanford Univer. Press, 1960.

THORNDIKE, E. L. *The fundamentals of learning.* New York: Teacher's College, 1932.

VOEKS, V. W. Acquisition of S-R connections: A test of Hull's and Guthrie's theories. *J. exp. Psychol.*, 1954, **47**, 137–147.

(Received August 23, 1960)

Psychological Review
1962, Vol. 69, No. 1, 54-61

THEORETICAL NOTES

TESTING THE NULL HYPOTHESIS AND THE STRATEGY AND TACTICS OF INVESTIGATING THEORETICAL MODELS

DAVID A. GRANT[1]

University of Wisconsin

Testing the null hypothesis, H_0, against alternatives, H_1, is well established and has a proper place in scientific research. However, this testing procedure, when it is routinely applied to comparing experimental outcomes with outcomes that are quantitatively predicted from a theoretical model, can have unintended results and bizarre implications. This paper first outlines three situations in which testing H_0 has conventionally been done by psychologists. In terms of the probable intentions or strategy of the experimenter testing H_0 turns out to be an appropriate tactic in the first situation, but it is inadequate in the second situation, and it is self-defeating with curious implications in the last situation. Alternatives to this conventional procedure are then presented along with the considerations which make the alternatives preferable to testing the usual H_0.

Three Applications of H_0 Testing

Probably the most common application of the tactic of testing H_0 arises when the independent variable has produced a sample difference or set of differences in the magnitude of the dependent variable. Quantitative predictions of the size of the difference or differences are not available. The experimenter wishes to know whether or not differences of the size obtained could have occurred by virtue of the operation of the innumerable nonexperimental factors conventionally designated as random. He sets up H_0 that the differences are zero; chooses a significance level, α; determines the set of hypotheses alternative to H_0 that he is willing to entertain, H_1; selects an appropriate test statistic, t, F, χ^2, U, T, or the like; and proceeds with the test. Rejection of H_0 permits him to assert, with a precisely defined risk of being wrong, that the obtained differences were not the product of chance variation. Failure of the test to permit rejection of H_0, which, unfortunately, is commonly termed "accepting" H_0, means that the obtained differences or greater ones would occur by chance with a probability greater than α. This situation is straightforward. The experimenter has limited aims. He has asked a simple question, and he has received a simple answer, subject only to those ambiguities which attend all experimental and inductive inference. His tactics are admirably suited to his strategic objective.

Another common but less satisfactory instance of testing H_0 arises when the results of pre-experimental matching or pretesting are to be evaluated. Here the experimenter has measured the dependent variable or some related variable before operation of the independent variable, and he devoutly hopes that the experimental and control groups are alike except for random differences. He is now relieved or chagrined, depending upon whether H_0 is "accepted" or "rejected" as a consequence of his test.

[1] The author is indebted to Arnold M. Binder of Indiana University whose arguments inspired him to make explicit some of the issues involved in using conventional analysis of variance procedures in testing the adequacy of a theoretical model. As this paper went through various revisions over a period of time the writer is correspondingly indebted to a number of supporting agencies: the Graduate Research Committee and the College of Letters and Science of the University of Wisconsin, the National Science Foundation, and finally to the Department of Psychology of the University of California, Los Angeles, his host during final preparation of the manuscript.

Even if H_0 is accepted his relief is tempered by some uneasiness. He knows that he has not proved, and indeed cannot prove, that H_0 is "true." His tactics in testing H_0 seem to be appropriate to the impossible strategic aim of proving the truth of H_0. Certainly, if he had a more reasonable aim he has adopted inappropriate tactics. Utilizing these tactics, the best he can do is to beat a strategic retreat, and if H_0 is accepted he can perhaps point out that he has used a very powerful test and that if there were real differences they were most likely very small. Although psychologists have never to my knowledge done so, he might be able to go one step further and point out that his testing procedure would reject H_0 a given percentage of the time, say, 90%, if the "true" difference had been as little as, say, one-tenth of an *SD*. This sort of statement of the power of a test is a commonplace in acceptance inspection (Grant, 1952, Ch. 13).

With the advent of more detailed mathematical models in psychology (e.g., Bush, Abelson, & Hyman, 1956; Bush & Estes, 1959; Goldberg, 1958; Kemeny, Snell, & Thompson, 1957) a new statistical testing situation is arising more and more frequently. The specificity of the predictions and perhaps the whole philosophy behind model construction pose a different kind of statistical problem than those faced by most psychological investigators in the past. It seems obvious that as the use of models becomes more widespread a greater number of investigators will face the problem of evaluating the correspondence between empirical data points and precise numerical predictions of these points. Unfortunately most of the procedures used to date in testing the adequacy of such theoretical predictions set rather bad examples. Probably the least adequate of these procedures has been that in which an H_0 of exact correspondence between theoretical and empirical points is tested against H_1 covering any discrepancy between predictions and experimental results.

Most models predict a considerable number of different aspects of the data, and some of these aspects are predicted with greater success than others (Bush & Estes, 1959, Chs. 14, 15, 17, 18). We shall restrict our discussion to the prediction of values along a curve which might be a learning curve. An idealized version of such a typical situation is presented in Figure 1. Here, the dependent variable, Y, is plotted on the vertical axis against the independent variable, X, on the horizontal axis. The theoretical model has led to an expression, $Y' = f(X)$, giving a set of k theoretical predictions, $Y'_1, Y'_2, \cdots, Y'_k$. The experiment has produced k empirical data points, a set of mean values, $\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_k$, corresponding to the values of the independent variable that were investigated, namely, $X_1, X_2, \cdots, X_k$. Individual observations tend to form normal distributions about each of the $\hat{Y}_i$, and these normal distributions tend to have equal σ's for all data points. In further discussion we shall assume that inaccuracies in the manipulation of the independent variable, X, can be ignored. The problem now is to investigate the goodness of fit of the Y'_i to the $\hat{Y}_i$ or the correspondence between the Y'_i and the $\hat{Y}_i$.

The tactics oriented toward accepting H_0 as corroborating the theory involve breaking down the jth individual obser-

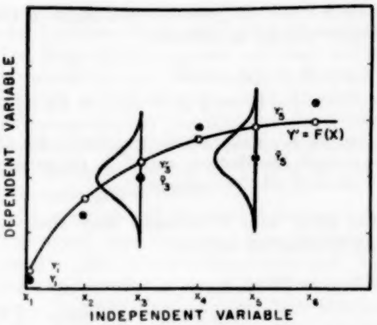

FIG. 1. Idealized situation involving the test of a theoretical function, $Y' = f(X)$. (Theoretical points, Y'_i, are represented by open circles; obtained means, $\hat{Y}_i$, are represented by solid circles.)

vation from the general mean of all of the observations, as follows:

$$Y_{ij} - \bar{Y} = (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - Y'_i) + (Y'_i - \bar{Y}) \quad [1]$$

where Y_{ij} is the jth observation in the ith normal distribution, and $\bar{Y}$ is the general mean of all observations.

the total sum of squares may then be partitioned as follows:

$$SS_{\text{Tot}} = SS_{\text{Dev Est}} + SS_{\text{Dev Theory}} + SS_{\text{Theory}} \quad [2]$$

where $SS_{\text{Dev Est}}$ is the sum of squares associated with the variation of individual measures from their means, $SS_{\text{Dev Theory}}$ is the sum of squares associated with the systematic departures of empirical data points from the theoretical points, and SS_{Theory} is the sum of squares associated with departures of the theoretical points from the general mean of the whole experiment.

If we suppose that the linear model for the analysis of variance holds, then:

$$Y_{ij} = \mu + T_i + D_i + e_{ij} \quad [3]$$

where μ is the population mean for all Y_{ij} over the specific values of the independent variable, X_i; T_i is the departure of the "true" theoretical value of Y'_i from μ; D_i is the discrepancy of the "true" value of $\bar{Y}_i$ from the true value of Y'_i; and e_{ij} is a random element from a normal distribution with a mean of zero and variance, σ_e^2, for all i.

For a fixed set of X_i the Ts and Ds may be defined so that $\sum T_i = \sum D_i = 0$. Under H_0 each $D_i = 0$. Under H_i some $D_i \neq 0$, and the variance of the D_i, $\sigma_D^2 \neq 0$. This last variance may be termed the true variance of the discrepancies from the theory over the particular set of X_i that was investigated.

The foregoing is a conventional analysis of variance model, and the F ratio of the $MS_{\text{Dev Theory}}$ divided by the $MS_{\text{Dev Est}}$ provides an excellent and powerful test of H_0 against H_1. The number of degrees of freedom for $SS_{\text{Dev Est}}$ will be $k(n-1)$ where n is the number of observations per data point, and the degrees of freedom for $SS_{\text{Dev Theory}}$ will be $k - n_T$, where n_T is the number of degrees of freedom lost in the process of fitting the model to the data. If this F is significant, we reject H_0, concluding that the discrepancies between the $\bar{Y}_i$ and the Y'_i are too great to be accounted for by the observed random variation in the experiment. In this conclusion we accept the 5% or 1% risk implied by our choice of α.

Logical difficulties arise when F fails of significance. H_0 remains tenable but is not proved to be correct. A tenable H_0 provides some support for the theory but only in the negative sense of failing to provide evidence that the theory is faulty. To assert that accepting the H_0 proves that the model provides a satisfactory fit to the data is an inaccurate and misleading statement. We may mean that we are satisfied, but others, especially proponents of other theories, will tend to regard our test as too lenient.

Failure to reject H_0, instead of producing closure, leaves certain annoying ambiguities, but the tactics of testing this particular H_0 imply a strategy that suffers from more serious defects that are readily apparent when the whole conception of testing a theory is carefully considered. To begin with, in view of our present psychological knowledge and the degree of refinement of available theoretical models it seems certain that even the best and most useful theories are not perfect. This means, in terms of the analysis of variance model, there will be some nonzero D_i's. H_0, then, is never really "true." Its "acceptance," rather than "proving" the theory, merely indicates that in this instance the D_i's were too small to be demonstrated by the sensitivity of the experiment in question. The tactics of accepting H_0 as proof and rejecting H_0 as disproof of a theory lead to the anomalous results that a small-scale, insensitive experiment will most often be interpreted as favoring a theory, whereas a large-scale, sensitive experiment will usually yield results opposed to the theory!

Curiously enough, even rejection of H_0 by means of a very stringent experimental test may be quite misleading as far as casting light on the adequacy of

the theory is concerned. If the D_i's are very small indeed the theoretical model may be a great improvement over anything else that is available and satisfactory for many purposes even though an extremely sensitive experiment were to reveal the nonzero D_i's. If our task, as scientists, were to test and accept or reject theories as they came off some assembly line the tactics of testing H_0 could be made in a satisfactory manner simply by requiring that the test be "sufficiently" stringent. In fact, our task and our intentions are usually different from testing products; what we are really up to resembles *quality control* rather than *acceptance inspection*, and statistical procedures suitable for the latter are rarely optimal for the former (Grant, 1952, Chs. 1, 13).

HYPOTHESIS TESTING VERSUS STATISTICAL ESTIMATION

An analogy will make clear the relation between testing tactics and the intention of the tester. Suppose that I wish to test a parachute; how should I go about it? How I should test depends upon my general intentions. If I want to sell the parachute and am testing it only to be able to claim that it has been tested, and I do not care what happens to the purchaser, then I should give the parachute a most lenient, nonanalytic test. If, however, I am testing the parachute to be sure of it for my own use, then I should subject it to a very stringent, nonanalytic test. But if I am in the competitive business of manufacturing and selling parachutes, then I should subject it to a searching, analytic test, designed to tell me as much as possible about the locus and cause of any failure in order that I may improve my product and gain a larger share of the parachute market. My contention is that the last situation is the one that is most analogous to that facing the theoretical scientist. He is not accepting or rejecting a finished theory; he is in the long-term business of constructing better versions of the theory. Progress depends upon improvement or providing superior alternatives, and improvement will ordinarily depend upon knowing just how good the model is and exactly where it seems to need alteration. The large D_i's designate the next point of attack in the continuing project of refining the existing model. Therefore attention should be focused upon the various discrepancies between prediction and outcome instead of on the over-all adequacy of the model.

In view of our long-term strategy of improving our theories, our statistical tactics can be greatly improved by shifting emphasis away from over-all hypothesis testing in the direction of statistical estimation. This always holds true when we are concerned with the actual size of one or more differences rather than simply in the existence of differences. For example, in the second instance of hypothesis testing cited at the beginning of this paper, where the investigator tests a pre-experimental difference, he would do better to obtain 95% or 99% confidence interval for the pre-experimental difference. If the interval is small and includes zero, he (and any other moderately sophisticated person) knows immediately that he is on fairly safe ground; but if the interval is large, even though it includes zero, it is immediately apparent that the situation is more uncertain. In both instances H_0 would have been accepted.

TESTING A REVISED H_0

Before turning to estimation procedures that are useful in examining the correspondence between experimental outcomes and predictions from a mathematical model, I shall digress briefly to outline a statistical testing method which can legitimately be used in appraising the fit of a model to data as shown in Figure 1. Basically the statistical argument in the proper test is reoriented so that rejection of H_0 constitutes evidence favoring the theory. The new H_0 is that the correlation between the predicted values, Y'_i, and the obtained values, $\bar{Y}_i$, is zero, after all correlation due to the fitting process has been eliminated. The alternative, H_1, against which H_0 is tested is that there is a correlation greater than zero be-

tween theoretical and empirical points. The four simple steps required to obtain the necessary F test are as follows:

1. Calculate $t_i = Y'_i - \bar{Y}$ for all i. $\Sigma t_i = 0$.[2]
2. Calculate $SS_{\text{Correspondence}} = n(\Sigma t_i \bar{Y}_i)^2/\Sigma t_i^2$, where n is the number of observations upon which each $\bar{Y}_i$ is based. Negative values of $(\Sigma t_i \bar{Y}_i)$ are treated as zero.
3. Obtain $MS_{\text{Correspondence}} = SS_{\text{Correspondence}}/n_T$, where n_T, the number of degrees of freedom involved in fitting the theoretical points to the empirical data, will ordinarily be the number of linearly independent fitting constants in the mathematical expression of the model.
4. Divide $MS_{\text{Correspondence}}$ by $MS_{\text{Dev Est}}$ to give $F_{\text{Correspondence}}$ which has n_T degrees of freedom for its numerator and $k(n-1)$ degrees of freedom for its denominator, k being the number of $\bar{Y}_i$. The test is one-tailed in the sense that negative values of $\Sigma t_i \bar{Y}_i$ are treated as zero values, so that the probability values of the F distribution must be halved, an unusual procedure with F tests in analysis of variance.

Following the above procedure, rejection of H_0 now means that there is more than random positive covariation between predicted and obtained values of the dependent variable.

This test is admirable in that it puts the burden of proof on the investigator, because a small-scale, insensitive experiment is unlikely to produce evidence favoring the model. Furthermore, if the model has any merit, the more sensitive the experiment, the more likely it is that a significant F, favoring the theory, will be obtained. Actually, the test is extremely sensitive to virtue in the theory, and therefore in the case of a moderately successful model and a moderately sensitive experiment both this F and the one testing the significance of systematic deviations from the model ($F = MS_{\text{Dev Theory}}/MS_{\text{Dev Est}}$) will tend to be significant. This outcome is no anomaly; it merely indicates that the model predicts some but not all of the systematic variation in the data. In short, progress is being made, but improvement is possible. The fact that simultaneous significance of both Fs, indicating general success and specific failures of a model, should be a commonplace points up the necessity of turning to methods of statistical estimation for a more adequate examination of the workings of a theoretical model.

Practical Estimation Methods for Investigation of Models

As is true of statistical tests, each method of statistical estimation has its advantages and limitations. In the investigation of the adequacy of theoretical curves in psychology there are reasons to believe that the simpler estimation methods have practical advantages over some of the more elegant procedures. To give a fairly complete view of the situation, methods of point and interval estimation of σ_D^2 and of the individual D_i will be described, and a brief evaluation of each method will be given.

Estimating σ_D^2. The variance of the discrepancies between the Y'_i and the $\bar{Y}_i$ condenses into a single number the adequacy of fit of the theoretical model. As such it is an excellent index for the evaluation of the model. The smaller the variance, σ_D^2, the better the model, and vice versa. As an estimate of the size of the discrepancies one might expect in future similar applications of the model, σ_D^2 is far more informative than any F test. Furthermore σ_D^2 is readily estimated in the case of homogeneity of the error variance, σ_e^2. The expected values of the relevant mean squares are as follows:

[2] In the unusual event where the general mean of the observations, $\bar{Y}$, is not used as a fitting constant for $Y' = f(X)$, the t_i must be computed as deviations from the mean of all the Y'_i, $\bar{Y}'$. The test will then be insensitive to discrepancies between $\bar{Y}'$ and $\bar{Y}$, and the interpretation will be somewhat equivocal. A separate test of H_0 that $\bar{Y}_{\text{population}}$ equals $\bar{Y}'$, is feasible, but here the experimenter is forced into the illicit posture of seeking to embrace H_0.

$$\text{Exp}(MS_{\text{Dev Theory}}) = \sigma_e^2 + n\sigma_D^2 \quad [4]$$

$$\text{Exp}(MS_{\text{Dev Est}}) = \sigma_e^2 \quad [5]$$

A maximum likelihood estimate of the variance of the discrepancies, $\hat{\sigma}_D^2$ is then:

$$\hat{\sigma}_D^2 = (MS_{\text{Dev Theory}} - MS_{\text{Dev Est}})/n \quad [6]$$

The accuracy of this estimator depends upon the number of degrees of freedom associated with $SS_{\text{Dev Theory}}$ and $SS_{\text{Dev Est}}$. The latter rarely poses any practical problem, but the former, in view of the predilection of psychologists for minimizing the number of data points, is quite critical. This is readily apparent when interval estimation of σ_D^2 is attempted.

Bross (1950) gives a convenient method for accurate approximation of the fiducial interval for σ_D^2, and in this case the fiducial and confidence intervals are essentially equal. The method will be outlined below for the 5% interval.

1. Obtain $\hat{\sigma}_D^2$ from Equation 6, above. (If the estimate is negative or zero, meaningful limits cannot be obtained.)
2. Find:

$$L = \frac{\dfrac{F}{F_{025(k-n_T,\ k[n-1])}} - 1}{\dfrac{F \cdot F_{025(k-n_T,\ \infty)}}{F_{025(k-n_T,\ k[n-1])}} - 1}$$

where:

$$F = MS_{\text{Dev Theory}}/MS_{\text{Dev Est}}$$

$F_{025(k-n_T,\ k[n-1])}$ is the entry in the 2.5% F table (Pearson & Hartley, 1954) for $n_1 = k - n_T$ and $n_2 = k[n-1]$; and $F_{025(k-n_T,\ \infty)}$ is the entry for $n_1 = k - n_T$ and $n_2 = \infty$.

3. Find:

$$L = \frac{F \cdot F_{025(k[n-1],\ k-n_T)} - 1}{\dfrac{F \cdot F_{025(k[n-1],\ k-n_T)}}{F_{025(\infty,\ k-n_T)}} - 1}$$

where $F_{025(k[n-1],\ k-n_T)}$ is the entry in the 2.5% F table for $n_1 = k[n-1]$, and $n_2 = k - n_T$; $F_{025(\infty,\ k-n_T)}$ is the entry for $n_1 = \infty$, and $n_2 = k - n_T$.

4. The upper and lower limits are then $\bar{L}\hat{\sigma}_D^2$ and $L\hat{\sigma}_D^2$, respectively. With less than 15–20 data points these limits will be found to be uncomfortably wide, a fact to bear in mind when designing an experimental test of a theoretical model. For example, in Figure 1, with 6 data points and two degrees of freedom for curve fitting, the limits might plausibly be 0–40, whereas with 14 data points the limits might be 0–12.

Aside from the considerable variability in the estimate of σ_D^2 which can be reduced by increasing the number of data points, there are two other important limitations to the use of estimates of the variance of the discrepancies in evaluating a model. First of all, the population value of σ_D^2 is completely dependent upon the particular values of the independent variable, X, which are chosen for the test of the model. Choice of two different sets of Xs could well lead to two entirely different values of σ_D^2, and both of these values could be perfectly accurate. Secondly, although σ_D^2 gives an over-all index of the adequacy of the model being tested, it condenses so much information into one measure that it does not permit pinpointing the especially large D_i's so that they can be given proper attention in considering revision of the model.

Estimating the D_i. The individual D_i may be estimated as points, or intervals may be established for the D_i, collectively or individually. As before, each method has its good points and its limitations.

Point estimation of the individual D_i's consists simply in comparing the individual data points, the $\bar{Y}_i$, with the fitted curve. It is a crude method, but it has served well in the past and represents the beginning of wisdom. For example, in Figure 1, the model builder might well note that the first three data points lie below the curve and ask himself if there is some special reason for this. He would also note that the greatest discrepancy occurs at $\bar{Y}_5$, where the neighboring discrepancies are in the other direction. The weakness of this simple method lies in the absence of a criterion which will assist the investigator in deciding which discrepancies should be

singled out for further attention and which may be disregarded because they are within the range of expected random variation. This defect is remedied by the interval estimation techniques.

Probably the ideal method of interval estimation is that in which intervals are established for the whole curve in one operation by finding the 95% confidence band. The method takes the theoretical curve as a point of departure, and the result is a pair of curves above and below the theoretical curve, which will tend in the case of random variation to contain between them 95% of the data points. Points lying outside the band are immediately suspect; they are the most promising candidates for attention in the next version of the model. There are two practical difficulties with this method. First, homogeneity of the error variance, σ_e^2, over all the X_i is required. And secondly, because errors in estimation of each fitting parameter must be taken into account, for all but the simplest curves (Cornell, 1956, pp. 184–186) the bands may be difficult[3] to obtain. Although the method is elegant, in practice it will rarely represent sufficient improvement over the final method, given below, to justify its use.

The last method seems to me to be the most useful and most robust and most flexible method. It can be widely applied, and the relative ease of application, coupled with its ability to discriminate between significant and random discrepancies make it superior to the other estimation methods. It also possesses the homely virtue of being readily understood. In contrast to the preceding method, this one takes as its point of departure the empirical means, and consists, simply, in computing the 95% confidence limits for each of the $\bar{Y}_i$. If there is homogeneity of variance, the error variance of each mean is taken simply as $\hat{\sigma}_e^2/n$; in cases of suspected heterogeneity, each mean must have its own estimate of error variance. This will, of course, be the variance of the distribution of Y_{ij} for each i, divided by n. When these limits have been obtained, attention is directed to instances where the theoretical curve lies outside the limits. In some cases, the investigator might choose to establish the 80% or 90% limits in order to direct his attention to less drastic departures of the experimental results from the model. Choice of an optimum level for the limits is hard to establish on a general a priori basis, but it is likely that limits narrower than the traditional 95% will be found more useful than the broader limits. Simple as this method is, it is hard to improve upon in actual practice. Instead of giving an almost meaningless over-all acceptance or rejection of a model, it directs attention to specific defects, its functioning improves as the precision of the experimental test is improved, and the investigator can set the confidence coefficient so as to increase its sensitivity to defect at a cost of a fairly well-specified percentage of false positives or wild goose chases. A final and often crucial advantage is that the confidence intervals, based as they are upon the experimental means, can be obtained in cases where the form of the theoretical function does not permit satisfactory estimation of its parameters, and the analysis of variance and confidence bands methods cannot properly be applied.

[3] A sufficient estimate of the error variance of each parameter must be available and independent of the estimates of all other parameters or else the covariances of all parametric estimates must be found and the theoretical function must have continuous first partial derivatives with respect to the parameters in order that the confidence bands may be found in the asymptotic case (Rao, 1952, pp. 207–208). Where an asymptote is involved in the fitting of the theoretical function, satisfactory independent estimators can rarely be obtained.

SUMMARY AND CONCLUSIONS

In this paper I have attempted to show that the traditional procedure of testing a null hypothesis (H_0) of a zero difference or set of zero differences is quite appropriate to the experimenter's intentions or scientific strategy when he is unable to predict differences of a specified size. When theory or other circumstances permit the prediction of differences of

specified size, using these predictions as the values in H_0 is tactically inappropriate, frustrating and self-defeating. This is particularly true when a theoretical curve has been predicted, and H_0 is framed in terms of zero discrepancies from the curve. If rejection of H_0 is interpreted as evidence against the theory, and "acceptance" of H_0 is interpreted as evidence favoring the theory, we find that the larger and more sensitive the experiment is, the more likely it will lead to results opposed to the theory; whereas the smaller and less sensitive the experiment, the more likely the results will favor the theory. Aside from this anomaly, which can be corrected by recasting H_0 in terms of a zero covariance between theoretical prediction and experimental outcome, hypothesis testing as a statistical tactic in this case implies an acceptance-inspection strategy. Acceptance-inspection properly involves examination of finished products with a view to accepting them if they are good enough and rejecting them if they are shoddy enough. The theoretician is not a purchaser but rather he is a producer of goods in a competitive market so that his examination of his theory should be from the standpoint of quality control. His idealized intentions are to detect and correct defects, if possible, so that he can produce a more adequate, more general theoretical model. Because his ideal strategy is not to prove or disprove a theory but rather to seek a better theory, his appropriate statistical tactics should be those involving estimation rather than hypothesis testing.

Examination of alternative techniques available for point or interval estimation of discrepancies between theoretical predictions and experimental outcomes or the over-all variance of these discrepancies suggests strongly that estimation of the confidence intervals for the means found along a theoretical curve is the most practical and most widely applicable general procedure. Other writers have recently emphasized the values of various estimation as opposed to hypothesis testing techniques (e.g., Bolles & Messick, 1958; Gaito, 1958; Savage, 1957) and it is hoped that considerations pointed out by them and points raised in this paper will be helpful to investigators who are in the process of examining theoretical models which lead to specific numerical predictions of experimental outcomes.

REFERENCES

BOLLES, R., & MESSICK, S. Statistical utility in experimental inference. *Psychol. Rep.*, 1958, **4**, 223–227.

BROSS, I. Fiducial intervals for variance components. *Biometrics*, 1950, **6**, 136–144.

BUSH, R. R., ABELSON, R. P., & HYMAN, R. *Mathematics for psychologists: Examples and problems.* New York: Social Science Research Council, 1956.

BUSH, R. R., & ESTES, W. K. (Eds.) *Studies in mathematical learning theory.* Stanford, Calif.: Stanford Univer. Press, 1959.

CORNELL, F. G. *The essentials of educational statistics.* New York: Wiley, 1956.

GAITO, J. The Bolles-Messick coefficient of utility. *Psychol. Rep.*, 1958, **4**, 595–598.

GOLDBERG, S. *Introduction to difference equations.* New York: Wiley, 1958.

GRANT, E. L. *Statistical quality control.* New York: McGraw-Hill, 1952.

KEMENY, J. G., SNELL, J. L., & THOMPSON, G. L. *Introduction to finite mathematics.* Englewood Cliffs, N. J.: Prentice-Hall, 1957.

PEARSON, E. S., & HARTLEY, H. O. (Eds.) *Biometrika tables for statisticians.* Cambridge, England: Cambridge Univer. Press, 1954.

RAO, C. R. *Advanced statistical methods in biometric research.* New York: Wiley, 1952.

SAVAGE, I. R. Nonparametric statistics. *J. Amer. Statist. Ass.*, 1957, **52**, 331–344.

(Received November 10, 1960)

Psychological Review
1962, Vol. 69, No. 1, 62–64

THE CONSTRUCTION OF *Q* SORTS: A CRITICISM

DONALD M. SUNDLAND[1]

Columbus Psychiatric Clinic

The widespread use of *Q* sorts makes it important that certain difficulties in item selection be emphasized. In this author's opinion, these difficulties are such as to suggest the discontinuance of the use of this technique for research purposes, unless corrective measures are devised.

The *Q* sort is a technique derived from *Q*-methodology practice (Mowrer, 1953; Stephenson, 1955). In using this technique, comparison is made between answers to a large number of "tests" (items) from different persons (usually judges), or from the same person from different points of view: "self, ideal self, self as seen by mother, etc." An implicit, underlying assumption is that the tests are not inherently correlated. Operationally speaking, if two judges agree that the quality "shy" is "very characteristic" of Person X, then their agreement about the qualities "bashful, modest, outgoing, extroverted, etc." is, at least partially, predetermined by the inherent correlations between the items. Now this would be all right if all the items were about shyness or nonshyness, then agreement between judges would mean that they agreed on the shyness of X. Characteristically this is not the case. Items relative to the broad range of personality traits are usually included in a *Q* sort. Under these circumstances the inclusion of items inherently related will lead to spurious correlations between *Q* sorts.

This difficulty has been pointed to before, but seemingly without appreciation of its seriousness. Stephenson (1955) wrote: "no set contained direct antonyms" (p. 257) suggesting that this was something to be avoided. Later Stephenson[2] said:

> you are quite right about . . . the use of direct antonyms. Little is to be gained by using words like introversive and extroversive in the same *Q*-sample, one of them is enough.

Cronbach (1953) said, in discussing criteria for the selection of items for *Q* sorts:

> statements, while logically bearing on the same domain, should represent a large number of continua. Correlating persons seems to have no advantage if we present items which all fall into one scale (p. 380).

Two types of statistical distortion are involved in this use of highly related items. The first is an inflation of the correlation coefficient, and the second is an inflation of the number of degrees of freedom. By using the total number of presumably unrelated items as the degrees of freedom, one markedly inflates the degrees of freedom in testing the significance of the correlation if many of the items are, in fact, correlated. A rather obvious example would be using only items related to a single dimension. If 30 items concerning a single trait were used to judge a person, the correlation between judges would be accurate and misleading at the same time. It would be accurate insofar as the judges' agreement about the trait, but it would be misleading about the person insofar as this is the only trait being considered. This is pretty obvious; however, now let us intermingle these 30 monotrait items among 70 other items, these 100 items supposedly representing the broad spectrum of personality traits. Further, let

[1] The author expresses his thanks to Robert J. Wherry, Edwin N. Barker, and Shephard Liverant for their valuable comments on this paper.

[2] W. Stephenson, personal communication, 1957.

us suppose that two judges have a zero correlation on these 70 other items, but they have a perfect, 1.00, correlation on the 30 monotrait items. The result of this intermingling would depend upon the placement of the 30, agreed upon items. The author devised a set of data to examine this problem. Seventy items were randomly distributed into the 11-category, approximately normal distribution shown in Table 1 and then minor adjustments were made to obtain an exactly zero correlation. The 30 agreed upon items were then intermingled in three ways. The first way was to place them in the extreme categories of the combined 100-item distribution shown in Table 1; the second was to place them in the middle categories; and the third was to distribute them proportionately. This procedure yields the correlations: 0.76, 0.11, and 0.30. A correlation of 0.26 for 100 items is significant at the 0.01 level. Considering the monotrait items as really one item, this example shows two types of overestimation: the correlations found range between .11 and .76 instead of from .00 to .06, and the degrees of freedom used are 98 (100–2) instead of 69 (71–2).

Let us look at another frequent use of the Q sort: self vs. ideal-self comparison. Where the monotrait items are rated in the same direction on both self and ideal-self Q sorts, the correlation will be overestimated. Where the monotrait items are rated in the opposite directions, the correlation will be underestimated. This distortion would be augmented if the experimenter were comparing Q sorts before and after therapy and if the monotrait items shifted their directions. For example, let us suppose that before therapy a patient sees himself as shy but as ideally not shy, and that after therapy he sees himself as not shy both actually and ideally. If the Q sort had many "shyness" items in it, the difference between the pretherapy and posttherapy correlation would be doubly overestimated since the pretherapy correlation would have been underestimated and the posttherapy correlation would have been overestimated.

TABLE 1

ITEM DISTRIBUTIONS

Number of Items	Categories										
	1	2	3	4	5	6	7	8	9	10	11
70 items	2	5	5	7	7	18	7	7	5	5	2
100 items	3	7	7	10	10	26	10	10	7	7	3

It would seem that a possible hope for the comparison of persons, or of different concepts of a person, would be the use of items whose interrelatedness is known. Levy and Dugan (1960) have proposed a method which offers a possible system for determining the interrelatedness of items. An investigator adapting their design would ask the subjects, from the same population as his experimental subjects, to describe a photograph with his Q-sort items. Each subject would be given a different photograph. The items could then be intercorrelated and factor analysed into orthogonal components. In using the Q sort with his experimental subjects the investigator could then obtain factor scores and use these for his comparisons. Since this is a somewhat laborious procedure it would seem to be desirable for investigators to come to some consensus about the items for two or three widely applicable Q sorts. Short of the availability of independent items or items with known dependencies, any kind of Q-sort study is subject to this distortion.

SUMMARY

It has been pointed out that due to the interrelatedness between items which constitute a Q sort, investigators obtain spurious results due to an overestimation of the correlation coefficient and of the degrees of freedom in their Q sort. The difficulty of determining the correct magnitude of the correlation coefficient and the correct number of degrees of freedom to be used indicates the desirability for

the development of alternate methods for Q-type studies.

REFERENCES

CRONBACH, L. J. Correlations between persons as a research goal. In O. H. Mowrer (Ed.), *Psychotherapy, theory and research.* New York: Ronald, 1953.

LEVY, L. H., & DUGAN, R. D. A constant error approach to the study of dimensions of social perception. *J. abnorm. soc. psychol.,* 1960, **61**, 21-24.

MOWRER, O. H. Q technique-description, history and critique. In O. H. Mowrer (Ed.), *Psychotherapy, theory and research.* New York: Ronald, 1953.

STEPHENSON, W. *The study of behavior.* Chicago: Univer. Chicago Press, 1955.

(Received December 1, 1960)

THE BRITISH JOURNAL OF PSYCHOLOGY

Edited by BORIS SEMEONOFF

Vol. 52, Part 4 November, 1961 20s. net

ZEITSCHRIFT FÜR EXPERIMENTELLE UND ANGEWANDTE PSYCHOLOGIE

Organ der Deutschen Gesellschaft für Psychologie

Band VIII/2